Exploring MicroRNA Strand Selection: Insights into Gene Regulation and Development

Author list

Dalton Meadows[1,2],

Hailee Hargis[1,2],

Amanda Ellis[1,2],

Marco Mangone[1,3]

Author affiliations

[1)] The Biodesign Institute at Arizona State University, 1001 S McAllister Ave, Tempe, AZ

[2)] School of Life Sciences, Arizona State University, 427 E Tyler Mall, Tempe, AZ

Author list footnotes

[3)] Lead contact

Corresponding author and e-mail address

Marco Mangone; mangone@asu.edu

Abstract

MicroRNAs (miRNAs) are important post-transcriptional regulators that play key roles in development, cell identity, and disease. While significant progress has been made in identifying miRNA targets, the mechanisms governing miRNA strand selection during biogenesis remain poorly understood. Here we investigate the hypothesis that miRNA strand selection is not random but follows specific thermodynamic and regulatory rules that vary across tissues and developmental stages.

Using *Caenorhabditis elegans* as a model system, we employed a novel high-throughput qPCR-based approach (HiTmiSS) to quantify miRNA strand selection throughout development and within somatic tissues, coupled with a novel machine learning model trained with our validated wet bench data, that can predict miRNA strand selection with high accuracy. When applied to vertebrates, our ML model has identified specific conserved features with important evolutionary implications for miRNA biogenesis. Our findings indicate that miRNA strand preference is widespread but varies dynamically, with a subset of miRNAs switching strand usage depending on developmental stage or tissue types. Furthermore, our analysis suggests that thermodynamic asymmetry alone does not fully explain strand selection, implicating additional cellular factors. This work provides a comprehensive framework for studying miRNA biogenesis, offering insights into regulatory principles governing strand selection and its broader implications in gene regulation, development, and disease across metazoans.

Introduction

miRNAs are short non-coding RNAs, ~22 nucleotides in length, that are incorporated into a large protein complex called the RNA-induced silencing complex (RISC), where they guide the interaction between the RISC and target mRNA by complementary base pairing primarily within 3' Untranslated (3'UTRs) regions in mature mRNAs (Bartel 2018). mRNAs targeted by the miRNA/RISC complex (miRISC) are typically held in translational repression. miRNAs and their 3'UTR targets are frequently conserved and play various roles in regulating fundamental biological processes across metazoans, such as development and morphogenesis.

miRNAs have undergone several pronounced expansions in metazoan genomes (Wolter et al. 2017). These expansions coincide with evolving organismal complexity yet rarely correspond with increases in protein-coding genes (Wolter et al. 2017). Eukaryotic genomes contain 100-2,000 distinct mature miRNAs in intergenic regions, introns, and operon-like clusters. Following transcription in long primary miRNA transcripts (pri-miRNAs) by the RNA polymerase II, they are processed in the nucleus by the microprocessor complex, comprising the RNase III enzyme Drosha and its cofactor DGCR8, into shorter hairpin structures known as precursor miRNAs (pre-miRNAs) (Bartel 2018). The two strands composing the pre-miRNAs are named the 5p and the 3p strand, respectively.

Pre-mRNAs are subsequently exported to the cytoplasm by Exportin-5, where they undergo further cleavage by the enzyme Dicer, producing mature miRNA duplexes (Bartel 2018). The Hsc70-Hsp90 chaperone complex facilitates the ATP-dependent

loading of the miRNA duplex into Argonaute, the protein that execute the silencing step. This complex is thought to induce conformational changes in Argonaute that enable duplex loading (Pare et al. 2009; Meijer et al. 2014; Nakanishi 2016). Once loaded, one strand of this duplex, the guide strand, is kept within the RISC, while the other, the passenger strand, is typically degraded. Both the 5p and the 3p strands can be selected for miRISC loading and act as guide strands. It is currently unclear how this process, termed 'miRNA strand selection,' is executed in detail, which proteins are involved in this decision, and how the guide strand is ultimately chosen between the two strands.

The mature miRNA within the RISC complex guides the complex to target mRNAs, leading to either translational repression or mRNA degradation, thereby influencing a wide array of cellular processes and contributing significantly to the regulation of gene expression. MiRNA biogenesis is a fundamental mechanism with far-reaching implications in development, homeostasis, and disease (Bartel 2018).

Recently, several groups (including ours) produced tissue-specific localization data for many miRNAs in *C. elegans* (Blazie et al. 2015; Blazie et al. 2017; Kotagama et al. 2019; Zhang et al. 2022; Schorr et al. 2023), flies (Carthew et al. 2017), mice (Isakova et al. 2020), rats (Smith et al. 2016), human tissues (Jima et al. 2010), and cancer (Ilieva et al. 2022). These results show that there are distinct functional miRNA populations in tissues, which can reshape transcriptomes and contribute to cell identity acquisition and maintenance. Since most of their targets are only predicted, how these events are initiated and maintained is still being determined.

miRNAs guide the RISC to complementary sites on target mRNAs by standard Watson-Crick and G:U wobble pairing. This pairing has been shown to require as little

as six consecutive nucleotides in the 5' end, or seed region, of the mature miRNA. Perfect complementarity between the miRNA seed region and its target is considered the canonical indicator of miRNA targeting. However, many recent studies indicate that miRNAs can target non-canonical elements in target mRNA (Wolter et al. 2017). Because these recognition elements are degenerate and small, each miRNA can potentially target hundreds of genes, dramatically reshaping cellular gene output. In this context, loading the correct miRNA strand into a RISC complex is extremely important since the 5p and the 3p target a completely different set of genes.

What mechanisms control Argonaute's choice in loading one strand of the miRNA duplex over the other? The predominant factors influencing miRNA strand selection are believed to be the specific 5' nucleotide sequences of each strand and the comparative thermodynamic stability exhibited by the two ends of the miRNA duplex (Medley et al. 2021). Studies performed using the human Ago2 protein, an Argonaute protein member of the RISC complex, show that the pocket responsible for binding the 5' nucleotide of the guide strand exhibits a strong preference for Uracil nucleotides (Frank et al. 2010).

This aligns with the prevalent occurrence of Uracil as the 5' nucleotide in miRNA guide strands (Wolter et al. 2017), whereas the passenger strand commonly initiates with a Cytosine at its 5' end.

Furthermore, the miRISC complex exhibits a preference for loading the end of the miRNA duplex with lower thermodynamic stability, although not all the time (Melendez and Levined August 24, 2009). The initial four nucleotides on each end of the duplex are believed to establish this thermodynamic asymmetry, and in some cases, a slight

difference in the number of hydrogen bonds can induce the preferential loading of the less stable strand *in vitro* (Meijer et al. 2014).

Consistently, the 5' ends of guide strands are enriched with purines (REF). In comparison, the 5' ends of passenger strands are predominantly pyrimidine-rich in both human and fly miRNAs, contributing to the thermodynamic asymmetry observed between the two ends of the duplex. In *C. elegans*, single amino acid substitutions within the MID (G553R) and PIWI (G716E, S895F, and S939F) domains of the Argonaute protein ALG-1 invert the strand selection process for specific miRNAs, suggesting a potential role for this protein in miRNA strand selection process (Zinovyeva et al. 2015). This indicates that while these mutations play a pivotal role, they do not consistently reverse strand selection for all miRNAs, suggesting that there may be additional internal or external factors that impact the selection of miRNA strand selection in *C. elegans*, depending on the specific miRNA involved. Tissue specific accessory factors could also be responsible of miRNA strand selection, altering strand choices based on tissue or developmental contexts, similarly to the role of the SR proteins in RNA splicing.

The nematode *C. elegans* represents an attractive system to study miRNA strand selection. This organism was the first multicellular system with an annotated 3'UTR reference dataset (3'UTRome), which was originally developed in 2011 within the modENCODE Project (Mangone et al. 2008; Mangone et al. 2010) and refined later on (Steber et al. 2019; Murari et al. 2024). The number of miRNAs is relatively small when compared to humans (253 miRNAs versus 1,917 in humans (miRbase)), and the size of 3'UTRs is also small (average size in *C. elegans* 230nt, ~1,5k in humans). In addition,

C. elegans has a complete cell-lineage map (Sulston 1983), its development is well studied at the physiological and molecular level (Chalfie et al. 1981), it has small, relatively simple tissues, and its transcriptome and 3'UTRome have been extensively characterized (Gerstein et al. 2010; Mangone et al. 2010; Jan et al. 2011; Ramani et al. 2011; Steber et al. 2019). *C. elegans* also possesses the best-annotated miRNA target predictions available for downstream studies (Bartel 2018). It has homologs for all members of the known human RISC complex (Grishok 2005) and is amenable to genetic manipulation.

After years of research, we have begun to understand the role of thermodynamics and nucleotide composition in this process, which can explain some but not all events. However, we still lack a general mechanistic framework that explains how miRNA strand selection is executed in eukaryotes.

Here, we developed a novel high-throughput approach to profile miRNA strand selection through *C. elegans* development and in somatic tissues. We have identified, pattern, motifs and trends that miRNA use in this process. These results were then used to train a machine learning model capable of accurately predicting strand usage with 83% accuracy and use it to perform a comparative analysis of miRNA strand selection between *C. elegans* and humans.

Materials and methods

All DI water was obtained from either a Purelab Option system (ELGA LabWater) or an Avidity Science system (Duo) and autoclaved before use. Chemicals were ordered

from Sigma Aldrich/Merck, VWR, or Thermo Fisher Scientific in molecular biology grade and stored according to company instructions. Gateway™ LR Clonase II Plus™ enzyme mixes, SuperScript™ III Reverse Transcriptase, and PowerTrack™ SYBR Green Master Mix were purchased from Invitrogen. Primers for the HiTmiSS assay were ordered from Life Technologies. Poly(U) Polymerase and restriction enzymes used in validation were ordered from New England Biolabs. The pCFJ150 plasmid was a gift from Erik Jorgensen. The OP50-1 *E. coli*, N2 *C. elegans* and EG6699 *C. elegans* strains were purchased through the Caenorhabditis Genetics Center.

Maintenance of *C. elegans* strains

All *C. elegans* strains were cultured on NGM plates seeded with OP50-1 bacteria. All transgenic lines produced for experiments described in this study were maintained at 24°C. When bacteria on each NGM plate had been cleared, worms were transferred onto new plates via washing into a 15 ml Falcon tube using 15 ml M9 media (22 mM KH_2PO_4, 42 mM Na_2HPO_4, 85.5 mM NaCl, 1 mM $MgSO_4$, sterilized by autoclave) and pelleted for 3 minutes at 400*g*. The M9 media was then aspirated to a volume of 500 ml and the animals were resuspended via gentle pipetting before being dropleted onto fresh, seeded NGM plates.

Synchronization of *C. elegans* strains

The EG6699 strain (Frokjaer-Jensen et al. 2008) was also maintained at 24°C. To synchronize animals for injection, confluent plates were bleached to isolate the embryos. Each plate was washed into a 15 ml Falcon tube with 15 ml DI water and pelleted for 3 minutes at 400*g*. The water was then aspirated and the animals were incubated in 5 ml of bleaching solution (0.5M NaOH, 10% bleach) for 20 minutes with gentle end-over-end rotation. After bleaching, the embryos were washed twice with DI water and dropleted onto fresh, seeded NGM plates.

For use in the HiTmiSS assay, both N2 and *pges-1::GFP::alg-1::unc54 3´UTR C. elegans* strains were synchronized as described above.

Primers design for the HiTmiSS Assay

To quantify microRNAs in high throughput, we designed 380 unique forward primers for use in the qPCR amplification step of the HiTmiSS assay. These primers were designed to be complimentary to either the 5p strand or the 3p strand of 190 most expressed microRNAs in the *C. elegans* genome, corresponding to 75% of all *C. elegans* miRNA deposited in miRbase (release 22.1). We did not consider microRNAs whose existence is not supported by broad experimental evidence nor any microRNAs predicted to exist. Few miRNAs were ignored because their corresponding primers had a very low Tm, making it impossible to process using our pipeline. All primers were ordered in four 96-well microplates from Life Technologies pre-diluted at a concentration of 50 mM (**Supplemental Table S1**). We designed a universal reverse primer complementary to the stem-loop sequence of the stem-loop poly(A) primer used to

prepare the cDNA for this qPCR reaction. 24 replicates of this primer were ordered from Life Technologies in a separate array, also at a concentration of 50 mM. Both forward and reverse primers were re-arrayed and diluted to a working concentration of 10 mM immediately prior to use.

***Singletons* and *Switchers* identification**

For each strand, we subtracted its relative quantity at one life stage versus the relative quantity of the complementary strand (i.e. for *let-7* 5p, we compared the abundance at embryo stage versus the abundance of *let-7* 3p at the embryo stage). This was repeated for each life stage at which the expression of that microRNA strand fell above our significance threshold, and any microRNAs for which their expression was greater at each such life stage were deemed *Singletons*. For microRNAs where the 5p strand was more highly quantified when significant at each life stage, these were classified as *5p Singletons*. For microRNAs where the 3p strand was more highly quantified when significant at each life stage, these were classified as *3p Singletons*. microRNAs which rose above our threshold of significance during at least one life stage, but whose expression levels did not display a consistent pattern of usage of one strand over the other were designated *Switchers*.

HiTmiSS Assay

To collect sufficient RNA for analysis of wild type microRNA, confluent NGM plates of N2 *C. elegans* were synchronized and either washed and pelleted immediately (embryo stage) or incubated at 24°C for 15 hours (larval 1 stage), 25 hours (larval 2 stage), 33 hours (larval 3 stage), 42 hours (larval 4 stage), or 51 hours (young adult stage) before washing and pelleting. All washes were performed in 15 ml Falcon tubes using 15 ml DI water and pellets were generated by centrifugation for 3 minutes at 400*g*, resulting in a final pellet volume of 500 µl to 1 ml. Sufficient worms to generate this size pellet differed by life stage, requiring 12 (larval 4 and young adult stages), 18 (larval 2 and larval 3 stages), or 24 (embryo and larval 1 stages) starting confluent plates.

This pellet was then resuspended to a volume of 1.5 ml in ice cold Triton X-100 lysis buffer (150 mM NaCl, 1.0% Triton X-100, 50 mM Tris-HCl pH 8.0, 0.5 mM EDTA) and transferred to a 2 ml Spectrum tube pre-filled with 0.5 mm zirconium beads. Using a Benchmark BeadBug D1030 microtube homogenizer, each tube was homogenized for 90 seconds, incubated for 2 minutes on ice, then homogenized for an additional 60 seconds. The resultant lysate mixture was centrifuged for 3 minutes using a VWR benchtop mini centrifuge and the liquid fraction was transferred into a 1.5 ml Eppendorf microcentrifuge tube. To minimize non-aqueous contaminants, this lysate was centrifuged again in the same manner and transferred to a 15 ml Falcon tube pre-chilled on ice.

To isolate total RNA, we used a DirectZol RNA MiniPrep Plus kit from Zymo Research, following kit instructions and performing the entire procedure at 4°C. Resultant RNA was eluted in 30 µL NF H_2O. To collect tissue-specific microRNA for

analysis of the intestine, we performed protein immunoprecipitation followed by microRNA isolation. The fluorescent *pges1::gfp::alg-1::unc54 3'UTR* strain (PMID: 31073019) of *C. elegans* previously prepared by our lab was grown to confluence, synchronized, and pelleted as described above for each life stage, concurrent with N2 *C. elegans* as a negative control.

Each pellet was resuspended to 1.5 μl using our Triton X-100 lysis buffer with added protease inhibitor (two Roche cOmplete ULTRA mini EDTA-free protease inhibitor tablets per 50 ml lysis buffer) and transferred to a zirconium bead tube to be homogenized as previously described. During this homogenization step, anti-GFP magnetic antibody beads (Chromotek GFP-Trap Magnetic Agarose) were prepared by diluting 25 μl bead slurry in 500 μl dilution buffer (150 mM NaCl, 50mM Tris-HCl pH 8.0, 0.5 mM EDTA) and rotating end-over-end at 4°C for 3 minutes.

The beads were pelleted using a benchtop centrifuge and buffer was aspirated using a magnetic tube rack. 500 μl dilution buffer were added once again, and this dilution process was repeated two more times. We then recovered aqueous lysate from the homogenized mixture as previously described and diluted it using 300 μl dilution buffer. This diluted lysate was transferred to a tube containing the prepared anti-GFP beads and left to rotate end-over-end at 4°C overnight. To clear unbound lysate, the beads were pelleted using a benchtop centrifuge and buffer was aspirated using a magnetic tube rack. 500 μl of wash buffer (150 mM NaCl, 50 mM Tris-HCl pH 8.0, 0.1% Triton X-100, 0.5 mM EDTA) was added, the beads were pelleted, and the buffer was aspirated using a magnetic tube rack. This wash step was repeated five additional times. The beads were then resuspended in 100 μl wash buffer and RNA was isolated using a

DirectZol RNA MiniPrep Plus kit from Zymo Research, following kit instructions and performing the entire procedure at 4°C. Resultant RNA was eluted in 30 μL NF H_2O.

For either whole worm or tissue-specific analysis, the remainder of the HiTmiSS protocol was the same. We prepared our cDNA from each life stage for qPCR analysis via polyuridinylation using Poly(U) Polymerase (NEB) as per manufacturer instructions followed by a 20-minute heat inactivation at 65°C. We performed a first-strand reaction on this polyuridinylated RNA using Superscript III Reverse Transcriptase and stem-loop poly(A) primers under recommended conditions (Invitrogen).

We then performed qPCR in high throughput using this cDNA, SYBR Green master mix (Invitrogen), and our primers as described above. We used Applied Biosystems Quantstudio 3 and Quantstudio 7 qPCR machines to perform the screen. For each life stage, four 96-well plates were prepared with 40 μL reactions in each well, along with a positive control reaction containing 10 pg pDONR 221 ROG instead of cDNA and the necessary forward and reverse primers to amplify *gfp*. Each plate was then re-arrayed in triplicate into a 384-well plate using a Biomek FX^P Laboratory Automation Workstation for a total of three 10 μL reactions per microRNA strand per life stage. An additional plate was prepared with no cDNA or plasmid spike-in DNA as a non-template control.

These qPCR reactions were carried out under standard SYBR Green reagent conditions on an Applied Biosystems QuantStudio 7. Relative microRNA levels were calculated using the $2^{-\Delta\Delta Ct}$ method as described above and normalized using the spike-in positive control. These results were then plotted logarithmically as a heatmap. Whole worm data were ordered by type (*5p Singleton*, *3p Singleton*, or *Switcher*), while

intestine data were clustered using average linkage. Both data were visualized using heatmapper.ca.

Cycle Threshold Analysis of the HiTmiSS Assay

The qPCR results containing the cycle threshold (Ct) values for each tested miRNA was exported in Microsoft Excel 2019 format. A comprehensive list of these data can be found in **Supplemental Table S2**. Ct values were averaged across experimental replicates, then subjected to a $2^{-\Delta\Delta Ct}$ analysis to calculate initial amounts of each microRNA in the sample, consistent with experimental standard (Livak and Schmittgen 2001). The Ct value for any given microRNA strand was subtracted from the Ct value of the non-template control (a qPCR reaction lacking any cDNA) for that microRNA strand. This was termed the ΔCt for that microRNA. The same was done for the positive control for that sample. Then for each sample this difference was subtracted from the difference obtained for that positive control, normalizing samples to the same positive control and giving us the $\Delta\Delta$Ct for that microRNA strand.

The inverse log base 2 of that $\Delta\Delta$Ct was then calculated to approximate the initial amount of that microRNA strand present in the cDNA relative to the positive control. This was deemed the $2^{-\Delta\Delta Ct}$ value for that microRNA strand.To determine whether a microRNA was considered present in a sample, we decided on a strict threshold of one standard deviation above the mean of all $2^{-\Delta\Delta Ct}$ values for that stage of development. MicroRNAs quantified at levels above this threshold were considered significant for the purposes of this study, while microRNAs not quantified at levels above this threshold

were not considered significant for the purposes of this study. This same process was repeated to determine significance for our intestine tissue-specific HiTmiSS assay.

qPCR of microRNAs targeting overexpressed intestinal 3′UTRs

To quantify the level of certain microRNA strands when the target 3′ UTRs of those strands are overexpressed in the intestine, we used two *C. elegans* strains previously published by our lab (Kotagama et al. 2019). Each expresses the promoter for intestine-specific gene *ges-1* fused to our lab's dual color ROG construct (a gene encoding *mCherry* and *gfp* separated by a trans-spliceable operon) fused to the 3′ UTR of either *fox-1* or *hrp-2*. Each of these strains was grown to confluence in 12 NGM plates in mixed stage and subjected to quantification using the same approach as the HiTmiSS assay described above, though rather than quantifying every microRNA in these animals, we quantified the 5p and 3p strand of *miR-54*, *miR-60*, *miR-232*, and *miR-239b*. These quantifications were carried out in 96-well plates using cDNA (both prepared from our strains as well as control cDNA gathered from N2 *C. elegans*), SYBR Green master mix, and the appropriate forward and universal reverse primers for the miRNAs of interest, rearrayed from our standard HiTmiSS plates.

Each 96-well plate contained two positive controls in *let-7* and *lin-4*, both of which are known highly expressed microRNAs in the intestine, as well as a non-template negative control for each set of primers. These qPCRs were carried out in experimental triplicate under standard SYBR Green reagent conditions on an Applied Biosystems QuantStudio 3.

Relative microRNA levels were calculated using the $2^{-\Delta\Delta Ct}$ method as described above and normalized using the average of *let-7* and *lin-4* expression for each sample. These results were then averaged across two replicates and analyzed using Microsoft Excel 2019. Data were visualized as $\log_2$ of the change in expression from wildtype N2 negative control to either of our experimental strains. The threshold of significance was set at 1 as per standard analytical procedure.

Structural Analysis of microRNAs

The experimentally validated *5p* and *3p Singletons* were obtained using our HiTmiSS assay. Each entry included the full hairpin sequence and its predicted secondary structure in dot-bracket notation (**Supplemental Table S2)**. To focus specifically on mature miRNA regions, only the sequences of each mature miRNA duplex was used for all downstream analyses. Since the *5p* and *3p Singleton* datasets contained different numbers of miRNAs, we balanced them prior to analysis by undersampling the larger group to match the sample size of the smaller class, ensuring unbiased statistical comparison between 5p and 3p strands.

We extracted 77 numerical features from each *5p* and *3p Singleton* (**Supplemental Table S3)**. The extracted features included: Base composition (fraction of A, U, G, C), motif presence (e.g. UGU, GAG, CCU), positional characteristics (first/last nucleotide identity), dinucleotide frequencies (e.g., di_GG, di_CC), GC content and GC-related asymmetry metrics (GC_diff, GC_ratio_5p3p), strand lengths, and number of

mismatches (unpaired positions in the dot-bracket structure). The full list of the features and their descriptions are shown in **Supplemental Table S3**.

Feature calculations were implemented in Python using regular expressions and sequence statistics functions. Structural mismatch counts were determined by counting unpaired characters within the uppercase-aligned mature duplex. Comparative statistical analysis between 5p and 3p strands was conducted using Welch's two-sample t-test. All features were tested independently, and results were ranked by p-value. Features with $p < 0.05$ were considered statistically significant. Bar plots were generated for each significant feature using the mean ± standard deviation across strands.

miRNA strand selection ML model

To predict dominant strand selection (5p vs. 3p) in miRNA hairpins, we utilized ChatGPT GPT-4o (OpenAI) to assist in developing and training a Random Forest classifier based on 77 biologically curated features, organized into functional categories capturing distinct structural and sequence-related aspects of miRNA hairpins. These features capture thermodynamic stability, base-pairing patterns, and positional relationships within the miRNA hairpin. The model was trained on our validated *5p* and *3p Singletons* identified using our HiTmiSS assay, and feature values were computed using custom Python scripts that parse dot-bracket RNA structures and extract both absolute and relative measures for each strand. The final model incorporates the following features divided into five categories. The first category is the *Dinucleotide*

Composition features, which quantify the frequency of specific two-nucleotide motifs within the 5p and 3p regions, capturing local sequence patterns that can influence strand selection and structural flexibility. The second category is the *First/Last Nucleotide Bias* features, which record the identity of the first and last nucleotides in each strand, reflecting known preferences in Argonaute loading where terminal nucleotides strongly impact strand choice. The third category is the *GC Content and Entropy* features, which assess the global nucleotide composition and thermodynamic diversity within each strand, including overall GC percentages, GC ratios between strands, and Shannon entropy measures that reflect structural variability and rigidity. The fourth category is the *Base-pairing Properties* feature, which describe the fraction of nucleotides engaged in base pairing within the 5p and 3p regions, providing a measure of local hairpin stability and asymmetry. Finally, the *General Structural Features*, which include the overall hairpin length and the length difference between the 5p and 3p strands, capturing broader geometrical properties that may indirectly affect miRNA maturation and strand retention (**Supplemental Figure S1**). The complete set of features, their description and the weight used are shown in **Supplemental Table S3**. The model showed high performance in distinguishing 5p and 3p usage from a pool of known *Singletons* as test dataset, (*5p Singleton*: accuracy 100%, precision 75%, recall 100%; *3p Singleton:* 82% accuracy, 100% precision, 82% recall) (**Supplemental Figure S2**).

ML model predictions in human miRNAs

Human microRNA (miRNA) expression data were obtained from a harmonized dataset from the human miRNA tissue atlas (Rishik et al. 2025) containing strand-specific Reads Per Million Mapped (RPMM) values across healthy human tissues. To restrict the analysis to non-diseased samples, we only considered entries where the Biotype column was labeled as ”tissue” or “EV”, “cell_line” (**Supplemental Table S6**), representing the average expression values across biological replicates of healthy tissues (total=337 datasets). For each miRNA duplex, we extracted the expression levels corresponding to the 5p and 3p strands, based on strand-specific annotations. We assigned strand dominance by summing RPMM values across all healthy tissue samples for the 5p and 3p strands separately. The strand with the higher total expression was designated as the dominant strand. We compared these dominat strands assignments with the predicted dominant strand assigned by our ML model. Each prediction was considered correct if the model-assigned strand matched the strand exhibiting greater expression across healthy tissues. Evaluation was performed on the topmost highly expressed 10, 50, 100 miRNAs ranked by the sum of 5p and 3p strands RPMM values. We also performed the analysis using the complete dataset (all).

Results

A novel HT qPCR approach to study miRNA biogenesis

miRNA identification from somatic tissues is challenging because miRNAs are small, not abundant, and have a short half-life. We tested several approaches and ultimately

developed a new, more robust HT qPCR-based approach capable of unambiguously identifying and quantifying miRNA expression and discriminating between 5p or 3p strand usage from total RNA isolated from somatic tissues. This method is a high throughput adaptation and general improvement of an existing methodology (Mei et al. 2012), which uses the polyuridine polymerase enzyme to elongate a polyU tail to all RNA transcripts in a given sample. This step extends the length of each miRNA present in the sample, which is then used as a template in reverse transcriptase reactions using a specially engineered universal stem-loop reverse primer (SL-polyA) (**Main Fig. 1A**).

To identify miRNA occurrence, measure their abundance, and pinpoint miRNA strand usage for each miRNA in the *C. elegans* genome, we adapted this approach to High-Throughput q-PCR experiments conducted in 96-well plate formats in triplicate. In this step, each well in the microplate contains forward primers complementary to either the 5p or 3p strand for mature *C. elegans* miRNAs, along with a universal reverse primer (**Main Fig. 1A**). We named this updated assay High-Throughput miRNA strand selection (HITmiSS) assay. HITmiSS enables the rapid and highly precise quantification of the entire miRNAome. It also accurately determines the abundance of the 5p or 3p strand for each miRNA present in the sample. (**Main Fig. 1A**).

miRNA expression is widespread throughout *C. elegans* development

We employed our HITmiSS methodology to systematically quantify miRNA occurrence and the usage of 5p and 3p strands in known miRNAs throughout *C. elegans* development. To achieve this, we extracted total RNA from *C. elegans* at six

developmental stages (embryos, larval stages 1–4, and young adults) and analyzed the miRNA transcriptome and strand selection for 190 miRNAs, representing 75% of all miRNAs annotated in miRbase, using a total of 380 unique forward primers across four plates (**Supplemental Table S1**).

We have performed 7,980 qPCR experiments independently, querying these 190 miRNAs throughout development in triplicate (1,140 experiments at each stage) (**Main Fig. 1B, Supplemental Fig. S3**). For each miRNA tested, we also repeated these experiments without adding the template cDNA to the reaction (negative control). We subtracted the signal obtained from our negative control from each miRNA tested, averaging the results for each triplicate experiment (see Materials and Methods). Using stringent filters, we detected the presence of 177 of these miRNAs (93.2%) in at least one developmental stage (**Supplemental Table S2**). The greatest number of miRNAs occur during L4 stage (81%), followed by young adult (75.5%), L3 (65%), L2 (56.8%), and lastly, Embryo and L1 (41.6% each) stages. This aligns with previous literature suggesting that many *C. elegans* miRNAs are tissue-specific, only becoming more abundant upon the development of those tissues later in the developmental cycle (Brosnan et al. 2021).

Consistent with past observations, most of the miRNAs detected in embryo and early larval stages play a functional role in embryonic lethality and early development. Namely, the *miR-35* and *miR-51* families, both important regulators of embryonic lethality, are the most highly expressed miRNAs during these early stages (Shaw et al. 2010). Similarly, the *let-7* and *lin-4* families, miRNAs that regulate adult phenotypes, were the most highly detected in L4 and young adult stages (Alvarez-Saavedra and

Horvitz 2010). Importantly, we did not detect significant expression of 13 miRNAs: *miR-791*, *miR-792*, *miR-1823*, *miR-2208a*, *miR-2209a*, *miR-2213*, *miR-2220*, *miR-2953*, *miR-4808*, *miR-4810b*, miR-*8201*, *miR-8203*, or *miR-8206*. Many of these miRNAs have not been experimentally validated, and they fall below our threshold.

In conclusion, gathering data using the HiTmiSS assay is robust, sensitive, reproducible, scalable and cost-effective. Although the filters we used in our analysis were very stringent, HiTmiSS was sensitive enough to detect more than 93% of all tested miRNAs above random background, including many which remain poorly characterized despite their relatively high expression.

Many miRNAs show consistent strand preference throughout development

Our method accurately quantifies miRNA abundance and detects which strand (5p or 3p) is present in each sample. Many miRNAs identified consistently use either the 5p or 3p strand as the guide strand, which we termed '*Singletons*', while others switch between 5p and 3p usage across different developmental stages. We named the second category '*Switchers*'. Importantly, the hits our assay produced align with past findings on the strand preference of well-characterized miRNAs such as *let-7, lin-4* (known *5p Singletons*), and the *miR-35* family (known *3p Singletons*) (Alvarez-Saavedra and Horvitz 2010). The screen produced hits for 177 miRNAs, which we divided into *5p Singletons*, *3p Singletons*, or *Switchers* (**Main Fig. 1B and Supplemental Table S2**).

The largest category identified is miRNA *Switchers*, with 88 hits (46.3%). These miRNAs do not consistently prefer one strand across all developmental stages and

instead switch strand usage between the 5p and the 3p strand throughout development. These included members of the well-characterized *miR-2* families and others, regulators of developmental timing, and heat shock (Alvarez-Saavedra and Horvitz 2010; Garcia-Segura et al. 2015). *Singletons* collectively represent the other half of the total hits obtained, and the strand usage is surprisingly not equally distributed, with *3p Singletons* accounting for nearly double the number of *5p Singletons* detected and are the second largest group detected, with 58 hits (30.5%) (**Main Fig. 1B**).

These included members of the *miR-35* and *miR-51* families, both regulators of embryonic lethality, members of the *miR-44* family, which regulate stress and starvation responses, as well as members of the *miR-58* family, responsible for locomotion, egg-laying, dauer entry, and body size (Alvarez-Saavedra and Horvitz 2010; Shaw et al. 2010; Garcia et al. 2014). Interestingly, some of the most highly detected *3p-singletons* are not well-defined functionally, such as *miR-60*, *miR-70*, and *miR-90*, all of which are known to broadly affect the stress response, though their specific impacts and targets remain unclear (Kagias et al. 2014; Garcia-Segura et al. 2015; Kato et al. 2016).

As mentioned, the least prevalent group detected is the *5p Singletons*, with 31 hits (16.3% of all tested miRNAs) corresponding to several well-characterized miRNA families (**Main Fig. 1B**). These included the *let-7* family and members of the *lin-4* family, both of which are known to regulate developmental timing and progression into the later stages of development (Alvarez-Saavedra and Horvitz 2010). We also detected members of the *miR-51* family, which regulates embryonic lethality, and the *miR-64* family, which regulates the heat shock response (Shaw et al. 2010; Kagias et al. 2014).

MiRNAs are often grouped by family containing similar seed regions and broadly

similar targets (Wolter et al. 2017). Some miRNA families use the same strand throughout development (*let-7*, *miR-64* family). However, the majority of them are spread between *5p Singletons*, *3p Singletons*, and *Switchers*, suggesting that even within the same family, with similar structure, use different rules to identify the used strand. For example, the *miR-51* family, a well-characterized developmental regulator (Brenner et al. 2012), has its members equally split between *5p Singletons* (*miR-51*, *miR-52*, *miR-53*) and *3p Singletons* (*miR-54*, *miR-55*, *miR-56*, *lsy-6*), with no *Switchers*.

To account for the strong strand preference observed among *Singletons*, we took the top 20 *5p Singletons* and the top 20 *3p Singletons* and compared the binding signature and nucleotide composition between the used strand versus the strand that gets discarded (**Main Fig. 1C, i and ii**). Interestingly, the pairing frequency does not vary between these miRNAs (**Main Fig. 1C,ii**), either in the 5p or 3p strand, with the exception of the last few nucleotides at the 3'end (**Main Fig 1C, I, red boxes**) which are commonly unpaired in the mature duplex.

When comparing the used vs. the discarded strand (**Main Fig. 1C, left panel vs right panel**), we detected no significant difference in the pairing signature at any position; however, the 5p strand, independently if is used or discarded, contains more unpaired nucleotides (**Main Fig. 1C, red boxes**). When narrowing our analysis to just the seed region (**Main Fig. 1C**, golden boxes), while we detect higher pairing than any other position within pre-miRNAs, we still do not detect any difference in binding frequency between the used and discarded strands, suggesting that other rules are also in place.

Next, we compared the nucleotide composition of these seed regions, hoping to

identify signatures that could explain the discrepancy in miRNA strand selection in *Singletons*. We determined the likelihood of any given nucleotide from positions 1 through 10 in each miRNA strand in the top 20 *Singletons* in each category (5p and 3p) (**Main Fig. 1C, ii**). We detected a highly conserved uracil nucleotide at position 1, as previously shown by us and others (Wolter et al. 2017), though this conservation dissolves from positions 2 through 10 (**Main Fig. 1C, ii**). Importantly, this uracil nucleotide is not equally conserved between *5p Singletons* and *3p Singletons*. 80% of *3p Singletons* start with this uracil nucleotide, while less than 40% occur in *5p Singletons*. This differential presence of this uracil may be caused by a non-stochastic pre-miRNA cleavage by the *C. elegans* DCR-1, which may possess a preference to cleave at this nucleotide. Unfortunately, the structure of DCR-1, including the cleavage site and nucleotide-binding pocket, remains unsolved, and more experiments need to be performed to validate if this cleavage preference occurs.

The intestine miRNA population is small but highly interconnected

In our experiments, we could not consistently detect 40.3% of the tested miRNAs throughout development. We believe that this is caused by either our stringent filters that exclude miRNAs with low signal in a specific developmental stage and/or because these miRNAs are not widely expressed in all stages and may be restricted in specific tissues. In addition, *Switchers*, which are the most abundant category in our whole worm screen, may not be 'true' *Switchers*, as they could act as either *3p Singletons* or *5p Singletons* in different somatic tissues and/or developmental contexts, skewing our

analysis.

Therefore, to better understand the rules behind miRNA strand selection, we decided to expand and deconvolve our analysis to somatic tissues throughout development. For these experiments, we focused on the intestine tissue because it is large, and its transcriptome has been extensively characterized by us and others (Haenni et al. 2012; Blazie et al. 2017).

We coupled our HiTmiSS assay with intestine-specific miRNA immunoprecipitation, as we have published in the past (Kotagama et al. 2019). More specifically, we forced the expression of the GFP-tagged *C. elegans* Argonaute ortholog *alg-1* in the intestine tissue using the intestine-specific promoter *ges-1p*, performed an α-GFP immunoprecipitation, recovered the miRNA strands loaded onto ALG-1, and used them as templates in our HiTmiSS assay (**Main Fig. 2A**). We performed these experiments in triplicates, focusing our analysis on the embryo and young adult stages for the most highly expressed miRNAs from our whole mixed tissue analysis.

We found that of the 96 miRNAs tested, 64 (68%) were present during at least one stage within the intestine tissue (**Main Fig. 2A-C**). Of these, 58 (91%) were *Singletons*, while 6 (9%) were *Switchers* (**Supplemental Fig. S4A**).

This strongly differs from our whole worm data, reinforcing that miRNAs previously labeled as *Switchers* may truly be *Singletons* on a tissue-by-tissue basis, only appearing as *Switchers* once all tissues are analyzed at once.

Interestingly, the proportion of *5p Singletons* to *3p Singletons* is not the same in the intestinal tissue as in the mixed tissue (**Supplemental Fig. S4A**). In our initial analysis, about two-thirds of *Singletons* were *3p Singletons*, yet in the intestine tissue, nearly two-

thirds of *Singletons* are *5p Singletons* (**Supplemental Fig. S4A**). This change implies additional factors influencing the strand selection decision specific to individual tissues, though further analyses in more tissues will be needed to confirm these trends.

Another surprise was the presence of intestine-specific *Switchers* (**Main Fig. 2A and Supplemental Fig. S4B**). This includes the miR-51 family, well-characterized regulators of embryonic lethality (Shaw et al. 2010). For these miRNAs, usage of both strands may be crucial to ensure proper development of intestine function during embryogenesis. In addition, we detected intestine tissue Switchers among the *miR-44, miR-58*, and *miR-71* families (**Supplemental Fig. S4B**). These miRNA families have diverse functions, from regulating developmental timing to starvation, hypoxia, stress response, and dauer entry (Kagias et al. 2014; Pandit et al. 2014; Garcia-Segura et al. 2015; Brunquell et al. 2017). Importantly, these are intestine-specific phenotypes, suggesting that strand switching in these miRNAs may have functional significance by expanding 3'UTR targeting repertoires in the intestine.

stage. These strains expressed the promoter for a single miRNA fused to GFP, and live imaging was used to validate the expression of these miRNAs at the appropriate life stage in the intestine (**Main Fig. 2C**). All four strains displayed the expected fluorescence.

miRNA target presence doesn't account for strand usage

We were concerned that the presence of miRNA targets in cells could stabilize a specific miRNA strand, thus influencing its detection by our HiTmiSS assay. In this

scenario, our approach might not accurately reflect genuine strand selection but rather detect artifacts resulting from target-induced stabilization. According to this model, the undetected strand lacks cellular targets, leading to its degradation and subsequent absence in our qPCR results.

To test this possibility, we overexpressed in the intestine tissue two 3'UTRs containing several miRNA binding sites for strands that are discarded as for our HiTmiSS results (*fox-1* 3'UTR and *hrp-2* 3'UTR) (**Supplemental Figure S5**). *Fox-1* 3'UTR contains binding sites for *miR-60 5p* and *hrp-2* contains dinging sites for *miR-232 5p* and *miR-239b 5p*). If this model was correct, we should observe an increase abundance of this typically discarded strand and a more equal ratio of 5p vs 3p usage. Our test did not detect any difference in miRNA strand usage for these miRNAs, which were always very similar to their wild type N2 controls, suggesting that the presence of a given miRNA target is not per se responsible for miRNA strand selection (**Supplemental Figure S5**). For example, *miR-60 3p*, which did not have binding sites to *fox-1* 3'UTR, remained the dominant detected strand in each experiment. (**Supplemental Figure S5**).

Taken together, these results indicate that our qPCR approach is not simply detecting stabilized miRNA strands due to the presence of targets. Rather, the relative strand abundance measured by the HiTmiSS assay likely reflects genuine strand selection, rather than artifacts introduced by target availability.

***5p Singletons* and *3p Singletons* follow different rules for selecting strand usage**

We next studied the structural features associated with Singletons, which could explain the strict and consistent strand usage within these miRNAs. When comparing the used and the discarded strand, the pairing frequency vary significantly within each *Singleton*, revealing new conserved patterns (**Main Fig. 3**). Specifically, the strand used in *Singletons* often has an unpaired nucleotide at position 12, while the discarded strand does not (**Main Fig. 3A**, yellow boxes). This lack of pairing in position 12 in both *5p Singletons* and *3p Singletons* usage is intriguing. In humans, certain Argonaute proteins perform cleavage of target RNAs between positions 10 and 11, which corresponds to positions 10 and 11 along the bound miRNA strand by Watson-Crick pairing (Nakanishi 2024). *C. elegans alg-1* is largely homologous to human Argonaute proteins, and it may retain this cleavage activity at similar target positions, leading to the degradation of the unused strand due to its Watson-Crick base pairing at this crucial position. In addition, we detected differences in the binding frequency of the first position nucleotide in both *5p Singletons* and *3p Singletons* (**Main Fig. 3A, blue boxes**). *3p Singletons* also exhibit unique pairings at positions 18-19 (**Main Fig. 3**, white box) not present in the discarded strain. These patterns are most clearly seen in specific *singleton* miRNAs, both *5p Singletons* like *miR-52* and *3p Singletons* such as *miR-75* (**Main Fig. 3B**).

An updated machine learning model for miRNA strand selection

We next sought to extend our findings by using our wet bench validated data as a training set for a machine learning model designed to predict 5p versus 3p strand usage in miRNA hairpins. We wanted to extract patterns and features that make *Singletons* so

unique and allow then to always use one strand over the other. We selected a Random Forest classifier, trained on experimentally validated *Singletons* identified through our HITmiSS assay. This model incorporates 77 features, systematically organized into five functional categories (dinucleotide composition, first/last nucleotide bias, GC content and entropy, base pairing properties, general structural features) that together capture critical aspects of miRNA mature duplex structure (**Supplemental Fig. S1–S2, Supplemental Table S3-S4, and Materials and Methods**). Dinucleotide Composition features capture the frequency of specific dinucleotide motifs within the 5p and 3p strands, which can influence local secondary structure and strand selection dynamics. First/Last Nucleotide Bias features record the identity of the first and last nucleotides of each strand, reflecting established preferences that impact Argonaute loading and strand stability. GC Content and Entropy features summarize the overall nucleotide composition and diversity of each strand, which assess rigidity, flexibility, and relative stability. Base-pairing Properties include the fraction of paired nucleotides in each arm, capturing local structural stability and asymmetry between the 5p and 3p strands. Finally, General Structural Features encompass broader geometrical characteristics, representing the overall size of the miRNA duplex and differences in strand lengths that may influence processing.Together, these features allow for accurate modeling of the structural and sequence-based factors governing strand selection. The model demonstrated strong performance in distinguishing 5p from 3p usage across a set of known miRNAs (**Supplemental Fig. S2, and Materials and Methods**).

***5p* and *3p Singletons* possess distinct structural features**

To explore the molecular determinants that differentiate strand selection, we conducted a comparative feature analysis of the mature *C. elegans* miRNAs.

We extracted and calculated all 77 features used in our ML model for the entire C, elegans miRNAome (374 miRNAs, miRbase) (**Supplemental Table S4**). We kept our definition of *Singletons* very stringent to increase the signal (confidence score >80%). We identified 54 *5p Singletons*, 52 *3p Singletons* and 268 *Switchers* (**Main Fig. 4A-B** and **Supplemental Table S4**). Importantly, many of the *Singletons* identified by the ML model were previously described as using only one strand (*let-7*, *lin-4*, etc.), reinforcing our confidence of the ML model.

Out of 77 extracted structural and sequence features, 13 exhibited statistically significant differences between the 5p and 3p strands ($p < 0.05$, Welch's t-test) (**Main Fig. 4B-E**). The most pronounced difference was observed in the identity of the first nucleotide of the 3p strand, which in *3p Singletons* was consistently a uracil ($p = 3.98 \times 10^{-}$) (**Main Fig. 4C**). Additionally, 3p strands showed an enrichment for specific motifs, particularly GAG ($p = 2.51 \times 10^{-}$), and were more likely to terminate in cytosine ($p = 5.69 \times 10^{-}$) (**Main Fig. 4E**). We also observed compositional differences: 3p strands displayed increased levels of cytosine at the first position and tended to be slightly shorter than their 5p counterparts ($p = 8.90 \times 10^{-3}$) (**Main Fig. 4D-E**).

These findings suggest intrinsic sequence and structural biases associated with strand identity that may reflect evolutionary pressures or mechanistic preferences in miRNA biogenesis. Together, these results highlight discrete, strand-specific features

that distinguish *5p* and *3p Singletons*.

Next, we decided to use our ML model and test if mutating these identified features, we could alter the overall miRNA structure and strand usage. In the case of *let-7*, altering the nucleotide pairing in position 11 and 12 (UP11-12) and/or altering the identity of the first nucleotide (5'U→C) significantly disrupted the strand selection and lowered its RNAfold predicted confidence as a *5p Singleton* up to 17%, supporting our model (**Main Figure 3C-D**). To further validate these findings, we altered the structure and composition of four other well-characterized *Singleton*: *miR-52, miR-75, lin-4, and miR-35* (**Supplemental Fig. S6**). In every case we saw a similar pattern that introducing bulges near the center of the mature hairpin disrupted predicted strand usage, while eliminating bulges in this region reinforced strand usage, further suggesting the importance of this region in our model of strand selection.

Next, we decided to use our ML model and test whether mutating the features identified in our *Singleton* data could alter overall miRNA structure and predicted strand usage. In the case of *let-7*, altering the nucleotide pairing in position 11 and 12 (UP11-12) and/or altering the identity of the first nucleotide (5'U→C) significantly disrupted the strand selection and lowered its RNAfold predicted confidence as a *5p Singleton* up to 17%, supporting our model (**Main Figure 3C-D**). To further validate these findings, we altered the structure and composition of four other well-characterized *Singletons*: *lin-4, miR-52, miR-75, and miR-60* (**Supplemental Fig. S6**). In every case we saw a similar pattern that introducing bulges near the center of the mature hairpin disrupted predicted strand usage, further reinforcing the importance of this region in our model of strand selection.

Similarly, the identity and pairing of the first nucleotide remained an important, though inconsistent indicator of predicted strand usage (**Supplemental Fig. S6**). Here we witnessed a stark difference in the magnitude of outcomes predicted as a result of these mutations between different miRNAs. While mutations of the first nucleotide in *5p Singletons* did not alter predicted strand usage more than 10%, similar mutations in 3p Singletons were predicted to change strand preference by as much as 30% (**Supplemental Fig. S6**). This recapitulates the trends in our HiTmiSS assay data and further reinforces our model of strand selection.

Strand Usage Bias Differs by Species

To investigate structural differences underlying miRNA biogenesis across species, we compared predicted hairpin features between *C. elegans* and human miRNAs using our trained model (**Main Fig. 5**) (**Supplemental Table S5**)**.** In *C. elegans*, strand usage was nearly balanced, with 51.3% of miRNAs predicted to favor the 5p strand and 48.7% favoring the 3p strand (192 vs. 182 predictions, respectively). This symmetry suggests that strand choice in *C. elegans* remains broadly permissive, possibly facilitating a more dynamic or context-specific regulatory output. In contrast, human miRNAs exhibited a marked asymmetry, with 75.9% of hairpins predicted to preferentially use the 5p strand (648 vs. 206 predictions) (**Main Fig. 5A**). In addition, statistical comparison revealed 13 features with significant differences ($p < 0.05$, Welch's t-test) between human and *C. elegans* miRNAs (**Main Fig. 5B**).

Among these, several features stood out with exceptionally strong statistical support. The most significant divergence was observed in guanine content on the 5p strand ($p = 5.62 \times 10^{-2}$), which was markedly higher in human miRNAs (**Main Fig. 5B**). Similarly, GG dinucleotide frequency ($p = 1.94 \times 10^{-2}$) was enriched in humans suggesting evolutionary divergence in sequence composition or processing preferences (**Main Fig. 5B**). Motif analysis further highlighted a strong enrichment of the CCU motif on the 3p strand in humans ($p = 9.37 \times 10^{-22}$), alongside a general increase in cytosine frequency in 3p regions (3p_C, $p = 1.56 \times 10^{-2}$). Additionally, di_CC ($p = 4.25 \times 10^{-2}$), another cytosine-related motif, was significantly more common in human sequences. Importantly these results are also evident in known miRNAs such as *let-7* (**Main Fig. 4B**). We have also applied our ML model to predict miRNA strand selection in additional species including mouse, rat, frog, zebrafish and fly (**Supplemental Table S5**).

Collectively, these findings suggest that human miRNAs exhibit cytosine- and dinucleotide-enriched motifs**,** especially in 3p strands**.** These compositional signatures likely reflect differences in strand selection, processing pathways, or functional targeting between the two species.

Our ML model can accurately predict human miRNA strand selection

To validate the accuracy of our machine learning model in predicting human miRNA strand selection, we compared model predictions to the comprehensive Human miRNA Tissue Atlas expression dataset in 65 healthy human organ systems (Rishik et al. 2025) (**Main Fig. 6**). When comparing model performance across subsets of highly expressed

miRNAs from this dataset, we observed a consistent trend: our model was very accurate when predicting strand usage among subsets of highly expressed miRNAs (**Main Fig. 6A**). Among the top 10 most highly expressed miRNAs, our model achieved an accuracy of 90%, surpassing even its accuracy in *C. elegans*. However, as we included more miRNAs from the atlas, the accuracy of the model, as expected, trended downward, reaching a lower yet still statistically significant ($p \approx 7.6 \times 10^{-}$□) accuracy of 59% when predicting strand usage among all human miRNAs in the dataset (**Main Fig. 6A**).

To refine our analysis and reduce potential noise from low-expression features, we focused on the top 50 most abundantly expressed miRNA duplexes (**Main Fig. 6B**). Among these, the model correctly predicted the dominant strand in 33 cases, corresponding to an overall accuracy of 66% ($p \approx 0.0164$) (**Main Fig. 6B**). Notably, our model displayed a strong bias toward the 5p strand, with several 3p dominant miRNAs being misclassified as *5p Singletons* (**Main Fig. 6B**). The 17 incorrect predictions were generally concentrated among miRNAs in the Atlas that had little difference in expression between their 5p and 3p strands. This in turn suggests that errors in our ML model predictions are more likely present when there is not a clear prevailing strand. Alternatively, since the difference in the expression levels of the 5p and 3p strand is so small, these miRNAs may be *Switchers* in Humans (**Supplemental Table S6**).

Of note, two of the mispredicted miRNAs, *miR-182* and *miR-378a,* have indeed been documented as *Switchers in vivo*. (Krishnan et al. 2013; Krist et al. 2015).

On the other hand, this misprediction of strand dominance often aligns with miRNAs where the confidence scores of the 3p and 5p strand are very close (+/- ~10%).

For example , *miR-199a* and *miR-199b*, both implicated in hypoxia responses and cardiac regulation (Rane et al. 2009), exhibited dominant expression of the 3p strand in healthy tissues but were predicted as 5p dominant by our ML model with a very small confidence of +/- ~7%. *miR-124*, a neuron-specific miRNA essential for neurogenesis, was similarly misclassified, despite showing strong 3p expression dominance in the dataset with a confidence of +/- 6% (Makeyev et al. 2007).

Still, the majority of miRNAs were correctly predicted across a broad range of organ systems, including several critical functional regulators known to exhibit strong strand bias, suggesting that our ML model trained in *C. elegans* Singletons is capable of accurately predict strand usage in vertebrates (**Main Fig. 6C and Supplemental Table S6**). For example, *miR-21*, one of the most abundant miRNAs across multiple organ systems and a well-established oncomiR (Rhim et al. 2022), was correctly predicted as a 5p dominant, with a greater than 50-fold difference in expression between its 5p and 3p strands. Similarly, *let-7a*, a central regulator of developmental timing and cell differentiation (Brueckner et al. 2007), was correctly classified as a 5p dominant with a 248-fold higher expression of the 5p strand relative to the 3p strand.

Other highly expressed miRNAs such as *miR-181a*, involved in T cell development (Kim et al. 2021), and *miR-100*, implicated in stem cell regulation (Pakravan et al. 2017), also exhibited strong 5p dominance and were accurately predicted by our ML model.

Overall, these findings demonstrate that the model is capable of reliably predicting strand usage for miRNAs in healthy human tissues, particularly in cases where strand bias is strong. Errors were enriched among duplexes with more balanced expression

levels or context-dependent strand usage, highlighting the complexity of predicting which strand will be used in cases of weak or dynamic regulation of strand selection.

Discussion

In this study, we present a comprehensive exploration of miRNA strand selection using a combination of high-throughput quantification, structural modeling, and predictive machine learning. Our goal was to study the molecular features that govern whether the 5p or 3p strand of a miRNA duplex is preferentially selected during biogenesis.

Using an in-house developed HiTmiSS assay, we performed the largest qPCR screen to date, performing ~8k qPCR experiments profiling miRNA strand usage throughout *C. elegans* development, and in the intestine tissue.

Our analysis led to the identification of experimentally validated *5p Singletons*, *3p Singletons* and *Switchers* throughout *C. elegans* developmental stages (**Main Fig. 1**).

We then used structural, thermodynamic and intrinsic properties from these experimentally validated data to develop a ML model capable of predicting high quality miRNA strand usage from a given pre-miRNA sequence and its corresponding hairpin structure. Furthermore, we extended our findings to vertebrate systems, comparing our ML model results to the Human miRNA Tissue Atlas.

We observed in our mixed tissue analysis in **Main Fig. 1** that a substantial portion of miRNAs act as *Switchers*, shifting strand usage across time, suggesting that strand

selection is not static but can vary by context. However, many well-characterized developmental regulators such as *let-7*, *lin-4* and *miR-35* are *Singletons*, highlighting the need of strand specificity in development where steady-state gene expression output is critical. In this context it is worth it to mention that *Singletons* in general are more abundant than *Switchers*, with even the most highly expressed *Switchers* detected at far lower levels than many *Singletons* (**Main Fig. 1B**). This is perhaps due to less stringent structural and thermodynamic control in these miRNAs.

Strand usage in the intestine, examined using tissue-specific ALG-1 immunoprecipitation (**Main Fig. 2**), revealed striking differences from whole-organism data. Here, *5p Singletons* outnumbered *3p Singletons*, inverting the developmental trend in our mixed tissue data. This suggests tissue-specific regulation of strand selection and supports the idea that miRNA strand choice is influenced not only by intrinsic structural features but also by the cellular environment and possibly tissue-specific protein cofactors.

Importantly, since we only tested one somatic tissue, we do not know whether any *Switchers* identified from our mix stage analysis are indeed tissue-specific *Singletons,* displaying opposing strand preference in different tissues and therefore appearing like *Switchers* when quantified in mixed tissue. More experiments in different somatic tissues need to be performed to address these questions.

One potential concern is that target presence may preferentially stabilize one strand of a miRNA duplex, thereby influencing its detection in our HiTmiSS assay. In this scenario, both 5p and 3p strands might initially be available for loading, but the

presence of a target could enhance the stability or retention of one strand, leading to an apparent, but potentially artifactual strand preference.

To test this, we overexpressed specific 3′UTRs known to contain binding sites for microRNA strands not typically abundant in the intestine (**Supplemental Figure S5**). Importantly, this manipulation did not alter the abundance of the corresponding strands detected by our HiTmiSS assay, indicating that our assay faithfully captures genuine strand usage rather than target-driven stabilization effects (**Supplemental Figure S5**).

When comparing the structural composition of each mature duplex, a series of patterns emerged. *5p Singletons* tend to have the first nucleotide unpaired to its corresponding strand in the duplex. This difference in pairing repeats in the middle of the duplex and drastically reverses toward the end (**Main Fig. 3A**). Contrarily, *3p Singletons* present a near-opposite trend, with the first nucleotide more often paired and the end portion of the sequence typically unpaired, though the center of the sequence shows similar unpairing to *5p Singletons*. Perhaps this highlights structural requirements that form the mechanistic basis to select for one strand over the other.

To address this question, we extracted 77 structural and compositional features from our validated *Singletons* and trained a Random Forest classifier (ML Model) to predict dominant strand usage. When comparing these features (**Main Fig 4**), we found several strand-specific differences, including a significantly higher occurrence of uracil at the first nucleotide in 3p strands and enrichment for GAG motifs.

We validated our predictions using *in silico* site-directed structural mutations in canonical miRNAs such as *let-7*, *lin-4*, and *miR-52*, *miR-75* (**Main Fig. 3C**, **Supplemental Fig. S4**). Disrupting features such as central bulges or first nucleotide

pairing, altered predicted strand preference to some degree, further confirming their role in the selection process. Surprisingly, the degree to which they impact miRNA strand preference was approximately equal, but when disrupted in tandem led to an even greater loss of predicted strand preference (**Main Fig. 3C**, **Supplemental Fig. S6**). In addition, these findings phenocopy the trends seen in our experimentally validated Singletons (**Main Fig. 3A**). This suggest, that the structural and thermodynamic features identified in our wet bench validated *5p* and *3p Singletons* are indeed critical regulator of strand selection in nematodes.

Expanding this framework, we then applied our ML model to predict strand usage across the human miRNA dataset (**Main Fig. 5A**). We were particularly curious to test if the structural and thermodynamics features and motifs identified in *C. elegans* were conserved in vertebrates.

Compared to the nearly equal distribution of 5p and 3p usage in *C. elegans*, human miRNAs showed a clear bias toward 5p strands, with over 75% of predictions falling in this category. This divergence likely reflects increased specialization in miRNA maturation and Argonaute loading in vertebrates, potentially driven by evolutionary selection for strand-specific targeting or stability. In this view, the reason for the 5p bias observed among human miRNAs may be due to the large expansion of miRNA families and miRNA hotspots seen across evolutionary times (Wolter et al. 2017), reinforcing strict rules which favor *Singletons* over *Switchers*.

We also found significant enrichment of guanine content in human 5p strands and cytosine-rich motifs in 3p strands (**Main Fig. 5B**). These compositional biases are

consistent with known differences in Argonaute protein specificity and may reflect lineage-specific constraints on strand functionality (Kotagama et al. 2024).

Importantly, these cross-species trends persisted in canonical miRNAs such as *let-7*, which was consistently predicted and validated as a *5p Singletons* across both systems, although at different confidence levels. This supports the robustness of our predictive ML model and highlights conserved features of miRNA processing, even in the presence of divergent strand preferences.

When we compare our ML model prediction in humans miRNA to the human miRNA Tissue Atlas, a public repository of known miRNA expression in their tissue of origin, an surprising trend emerges. Our ML model developed in *C. elegans* is capable of accurately predicting miRNA strand usage in humans (**Main Fig. 6**). This is unexpected, considering the vast evolutionary gap (~600 million years) between these two systems, which share less than 35% of nucleotide conservation in genes (Stein et al. 2001). Despite this, the features we have identified seems to be evolutionary conserved.

This suggests that miRNA strand selection is a fundamental cellular process which has evolved very little over time, even as miRNA regulatory networks expanded in complexity with the increased cellular specialization. This becomes more evident when we overlayed miRNA strand usage in human tissues of different origins. Here, the same miRNA always tends to use the same strand in different somatic tissues.

In conclusion, our findings illustrate the complexity of miRNA strand selection and reveal that this process is not governed solely by thermodynamic asymmetry. Instead, a combination of structural motifs, sequence biases, and cellular context dictate strand usage. Our work establishes a predictive framework for understanding strand selection

but also opens new avenues for studying the evolution of miRNA regulatory logic across metazoans.

Data Availability

The raw data produced by the HiTmiSS assay are included in the **Supplemental Table S2**. The ML model is included as Supplemental Data and is also available upon request.

Funding

This work was funded by the National Institutes of Health 5R01GM118796 awarded to M.M.

Acknowledgments

We thank Dr. Murugan for assistance in HiTmiSS assay optimization.

Author Contributions

D.M. and M.M. designed the experiments. D.M. performed the experiments. D.M. and M.M. prepared the ML model and performed the bioinformatics analysis. D.M., A.E. and H.H. performed the HiTmiSS assay experiments. D.M. and M.M. prepared the main and supplemental figures and supplemental tables. D.M. and M.M. wrote the manuscript. All authors have read and approved the manuscript.

Declaration of Interests

The authors have no conflicts of interest.

Figure titles and legends

Main Figure 1: HiTmiSS Assay and Analysis on Mixed Tissue. **(A)** HiTmiSS protocol for RNA extraction, extension, reverse transcription, and quantification in high throughput. Steps include (1) *C. elegans* growth and homogenization, (2) RNA extraction and isolation, (3) RNA polyuridinylation, extension, and reverse transcription (details shown in ii), (4) addition of miRNA-strand-specific forward primers, (5) loading into 384-well plates, and (6) quantification. **(B)** Heatmap of miRNA expression (n=3) in mixed tissue clustered by quantity, strand use, and developmental stage. Stages include embryo (E), larval 1 through 4 (L1-L4) and young adult (YA). The heatmap legend represents $\log_{10}$ of expression values for each miRNA strand normalized by spike-in GFP control at each stage from -4 (deep purple) to 4 (bright yellow). Black color indicates instances of data which fall below strict expression threshold (one standard deviation above mean of all expression values at that stage). **(C)** Heatmap of pairing frequency by position for the used (left) and discarded (right) strand of the 20 most highly quantified *5p* and *3p Singletons* in our mixed tissue dataset. Average pairing frequency and logo plot of nucleotide frequency at each position displayed below.

Main Figure 2: Analysis of Intestine Tissue-Specific miRNAs. **(A)** Protein immunoprecipitation of GFP-tagged ALG-1 under the control of the intestine-specific promoter *ges-1p* for subsequent RNA isolation and quantification using the HiTmiSS assay. **(B)** Heatmap of miRNA expression in intestine tissue clustered by strand and developmental stage. The heatmap legend represents $\log_{10}$ of expression values for each miRNA strand normalized by spike-in GFP control at each stage from -4 (deep purple) to 4 (bright yellow). Black color indicates instances of data which fall below strict expression threshold (one standard deviation above mean of all expression values at that stage). **(C)** Representative brightfield (left) and GFP (right) images of *C. elegans* strains expressing GFP under the control of promoters for miRNA known to be highly expressed in the intestine, either during embryo stage (top two) or young adult stage (bottom two).

Main Figure 3: Structural and Thermodynamic Determinants of Strand Selection. **(A)** Pairing frequency by position of the used (green) and discarded (red) strand for the most highly quantified *5p Singletons* and *3p Singletons* in mixed tissue (n=20). Positions at which pairing frequency most differs are boxed, including the first nucleotide (light blue), center of the duplex (gold), and end of the strand (white). **(B)** Representative pri-miRNA hairpins for a *5p Singleton* (*miR-52*) and a *3p Singleton* (*miR-75*) demonstrating occurrence of trends in pairing as described above. Mature duplex regions highlighted in pink. **(C)** Changes in *Singleton* predictive confidence for the miRNA *let-7* as a result of *in silico* site-directed mutations. RNAfold secondary structures of pri-miRNA with mature 5p strand (blue) and mature 3p strand (red) from left to right: wildtype (WT),

unpaired nucleotides at positions 11 and 12 via GG→UU substitution (UP11-12), first nucleotide U→C substitution (1U → C), combination of both prior mutations (Combined). Table shown displays probability that each mutation will yield a *5p Singleton* (5p) or *3p Singleton* (3p). **(D)** Bar chart displaying probability as described above.

Main Figure 4: Analysis of Mature miRNA Duplex Features in *C. elegans*. (A) Frequency distribution of predicted *5p Singleton* confidence across all miRNAs in *C. elegans*. *3p Singletons* (yellow) are defined as miRNAs falling below a predicted 5p confidence of 20% or below; *Switchers* are defined as miRNAs with a predicted 5p confidence between 20% and 80%; *5p Singletons* (red) are defined as miRNAs with a predicted 5p confidence of 80% or greater. **(B)** Predicted categorization of all *C. elegans* miRNAs by our model as per the above threshold. **(C)** Predicted categorization of *C. elegans* miRNAs for which uracil is the first nucleotide of the 3p strand. **(D)** Length difference (5p strand - 3p strand) for each category of *C. elegans* miRNAs. *5p Singletons* display significantly ($p<0.05$) longer 5p strands than 3p strands when compared either to *3p Singletons* or to *Switchers*. **(E)** Heatmap of 32 features including GC content, miRNA length, hydrogen bonding, nucleotide pairing, strand stability, and MFOLD-derived parameters enrichment in *5p* and *3p Singletons*. Heatmap values represent mean values of each feature across the miRNA categories. For binary features (i.e. 5p_has_AAU, which is either 0 or 1) the number represents the proportion (e.g. 0.3 = 30% of miRNAs in that category have an AAU motif in the 5p strand). For continuous features (i.e. gc_content, pairing_frac_5p), the number is the average value

across all miRNAs in that category. For nucleotide composition features (i.e. 5p_A, 3p_U), the number reflects the fraction of that nucleotide within the strand. Features clustered by average linkage.

Main Figure 5: Comparative Analysis of Mature miRNA Duplex Features between *C. elegans* and *H. sapiens*. (A) Frequency distribution of predicted *5p Singleton* confidence across all miRNAs in *H. sapiens*. *3p Singletons* (yellow) are defined as miRNAs falling below a predicted 5p confidence of 20% or below; *Switchers* are defined as miRNAs with a predicted 5p confidence between 20% and 80%; *5p Singletons* (red) are defined as miRNAs with a predicted 5p confidence of 80% or greater. **(B)** Four most significant ($p<0.05$) structural differences between *H. sapiens* (left) and *C. elegans* (right) mature miRNA duplexes. From top left to bottom right: mean occurrence of guanine in the 5p strand (5p_G), mean occurrence of guanine-guanine dinucleotide in either mature strand (di_GG), mean occurrence of cytosine-cytosine dinucleotide in either mature strand (di_CC), binary classifier for presence of cytosine-cytosine-uracil trinucleotide in the 3p strand (3p_has_CCU). **(C)** Alignment of *C. elegans* (*cel*) *let-7* and five *H. sapiens* (*hsa*) *let-7* homologs to demonstrate differences in CCU occurrence within the 3p strand (red boxes). **(D)** Heatmap of differences in mean value for each of the 77 model features between each category of miRNA in *C. elegans* (c) and *H. sapiens* (h). For binary features (i.e. 3p_has_CCU, which is either 0 or 1) the number represents the proportion (e.g. 0.3 = 30% of miRNAs in that category have an CCU motif in the 3p strand). For continuous features (i.e. gc_content, au_content), the number is the average value across all miRNAs in that category. For nucleotide

composition features (i.e. 5p_A, 5p_U), the number reflects the fraction of that nucleotide within the strand. Features clustered by average linkage.

Main Figure 6: **Model Accuracy for the Human miRNA Tissue Atlas Dataset**. **(A)** Model accuracy expressed as a percentage of correct predictions for four different subsets of *H. sapiens* miRNAs (total 854 mature duplexes). From left to right, ranked by total expression of mature 5p and 3p strand across all tissue types: predicted accuracy among top 10 (9 correct, 1 incorrect), top 50 (33 correct, 17 incorrect), top 100 (66 correct, 34 incorrect), and all (502 correct, 352 incorrect) miRNAs in this dataset. **(B)** Confusion matrix of the strand usage prediction model among the top 50 most highly expressed *H. sapiens* miRNAs. Summarizes differences in 5p and 3p strand usage between experimentally validated (miRNA Atlas) and model-derived (ML Model) data. **(C)** Heatmap of prediction accuracy, either correct (red) or incorrect (blue) for the top 50 most highly expressed *H. sapiens* miRNAs across 65 *H. sapiens* tissue types and organs. Data are clustered using average linkage.

Supplemental Information titles and legends

Supplemental Table S1: **Primers used in this study.**

Supplemental Table S2: **HiTmiSS assay results.**

Supplemental Table S3: **Features used in our ML model and their weights.**

Supplemental Table S4: **ML prediction in *C. elegans*.**

Supplemental Table S5: **ML prediction in vertebrates.**

Supplemental Table S6: **Comparison of ML model with miRNA tissue Atlas.**

References

Alvarez-Saavedra E, Horvitz HR. 2010. Many families of C. elegans microRNAs are not essential for development or viability. *Curr Biol* **20**: 367-373.

Bartel DP. 2018. Metazoan MicroRNAs. *Cell* **173**: 20-51.

Blazie SM, Babb C, Wilky H, Rawls A, Park JG, Mangone M. 2015. Comparative RNA-Seq analysis reveals pervasive tissue-specific alternative polyadenylation in Caenorhabditis elegans intestine and muscles. *BMC Biol* **13**: 4.

Blazie SM, Geissel HC, Wilky H, Joshi R, Newbern J, Mangone M. 2017. Alternative Polyadenylation Directs Tissue-Specific miRNA Targeting in Caenorhabditis elegans Somatic Tissues. *Genetics* **206**: 757-774.

Brenner JL, Kemp BJ, Abbott AL. 2012. The mir-51 family of microRNAs functions in diverse regulatory pathways in Caenorhabditis elegans. *PloS one* **7**: e37185.

Brosnan CA, Palmer AJ, Zuryn S. 2021. Cell-type-specific profiling of loaded miRNAs from Caenorhabditis elegans reveals spatial and temporal flexibility in Argonaute loading. *Nat Commun* **12**: 2194.

Brueckner B, Stresemann C, Kuner R, Mund C, Musch T, Meister M, Sultmann H, Lyko F. 2007. The human let-7a-3 locus contains an epigenetically regulated microRNA gene with oncogenic function. *Cancer Res* **67**: 1419-1423.

Brunquell J, Snyder A, Cheng F, Westerheide SD. 2017. HSF-1 is a regulator of miRNA expression in Caenorhabditis elegans. *PloS one* **12**: e0183445.

Carthew RW, Agbu P, Giri R. 2017. MicroRNA function in Drosophila melanogaster. *Semin Cell Dev Biol* **65**: 29-37.

Chalfie M, Horvitz HR, Sulston JE. 1981. Mutations that lead to reiterations in the cell lineages of C. elegans. *Cell* **24**: 59-69.

Frank F, Sonenberg N, Nagar B. 2010. Structural basis for 5'-nucleotide base-specific recognition of guide RNA by human AGO2. *Nature* **465**: 818-822.

Frokjaer-Jensen C, Davis MW, Hopkins CE, Newman BJ, Thummel JM, Olesen SP, Grunnet M, Jorgensen EM. 2008. Single-copy insertion of transgenes in Caenorhabditis elegans. *Nat Genet* **40**: 1375-1383.

Garcia SM, Tabach Y, Lourenço GF, Armakola M, Ruvkun G. 2014. Identification of genes in toxicity pathways of trinucleotide-repeat RNA in C. elegans. *Nature structural & molecular biology* **21**: 712-720.

Garcia-Segura L, Abreu-Goodger C, Hernandez-Mendoza A, Dimitrova Dinkova TD, Padilla-Noriega L, Perez-Andrade ME, Miranda-Rios J. 2015. High-Throughput Profiling of Caenorhabditis elegans Starvation-Responsive microRNAs. *PloS one* **10**: e0142262.

Gerstein MB Lu ZJ Van Nostrand EL Cheng C Arshinoff BI Liu T Yip KY Robilotto R Rechtsteiner A Ikegami K et al. 2010. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science* **330**: 1775-1787.

Grishok A. 2005. RNAi mechanisms in Caenorhabditis elegans. *FEBS Lett* **579**: 5932-5939.

Haenni S, Ji Z, Hoque M, Rust N, Sharpe H, Eberhard R, Browne C, Hengartner MO, Mellor J, Tian B et al. 2012. Analysis of C. elegans intestinal gene expression and polyadenylation

by fluorescence-activated nuclei sorting and 3'-end-seq. *Nucleic Acids Res* **40**: 6304-6318.

Ilieva M, Panella R, Uchida S. 2022. MicroRNAs in Cancer and Cardiovascular Disease. *Cells* **11**.

Isakova A, Fehlmann T, Keller A, Quake SR. 2020. A mouse tissue atlas of small noncoding RNA. *Proc Natl Acad Sci U S A* **117**: 25634-25645.

Jan CH, Friedman RC, Ruby JG, Bartel DP. 2011. Formation, regulation and evolution of Caenorhabditis elegans 3'UTRs. *Nature* **469**: 97-101.

Jima DD, Zhang J, Jacobs C, Richards KL, Dunphy CH, Choi WW, Au WY, Srivastava G, Czader MB, Rizzieri DA et al. 2010. Deep sequencing of the small RNA transcriptome of normal and malignant human B cells identifies hundreds of novel microRNAs. *Blood* **116**: e118-127.

Kagias K, Podolska A, Pocock R. 2014. Reliable reference miRNAs for quantitative gene expression analysis of stress responses in Caenorhabditis elegans. *BMC Genomics* **15**: 222.

Kato M, Kashem MA, Cheng C. 2016. An intestinal microRNA modulates the homeostatic adaptation to chronic oxidative stress in C. elegans. *Aging (Albany NY)* **8**: 1979-2005.

Kim C, Ye Z, Weyand CM, Goronzy JJ. 2021. miR-181a-regulated pathways in T-cell differentiation and aging. *Immun Ageing* **18**: 28.

Kotagama K, Grimme AL, Braviner L, Yang B, Sakhawala RM, Yu G, Benner LK, Joshua-Tor L, McJunkin K. 2024. Catalytic residues of microRNA Argonautes play a modest role in microRNA star strand destabilization in C. elegans. *Nucleic Acids Res* **52**: 4985-5001.

Kotagama K, Schorr AL, Steber HS, Mangone M. 2019. ALG-1 Influences Accurate mRNA Splicing Patterns in the Caenorhabditis elegans Intestine and Body Muscle Tissues by Modulating Splicing Factor Activities. *Genetics* **212**: 931-951.

Krishnan K, Steptoe AL, Martin HC, Wani S, Nones K, Waddell N, Mariasegaram M, Simpson PT, Lakhani SR, Gabrielli B et al. 2013. MicroRNA-182-5p targets a network of genes involved in DNA repair. *RNA* **19**: 230-242.

Krist B, Florczyk U, Pietraszek-Gremplewicz K, Jozkowicz A, Dulak J. 2015. The Role of miR-378a in Metabolism, Angiogenesis, and Muscle Biology. *Int J Endocrinol* **2015**: 281756.

Livak KJ, Schmittgen TD. 2001. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* **25**: 402-408.

Makeyev EV, Zhang J, Carrasco MA, Maniatis T. 2007. The MicroRNA miR-124 promotes neuronal differentiation by triggering brain-specific alternative pre-mRNA splicing. *Mol Cell* **27**: 435-448.

Mangone M, Macmenamin P, Zegar C, Piano F, Gunsalus KC. 2008. UTRome.org: a platform for 3'UTR biology in C. elegans. *Nucleic Acids Res* **36**: D57-62.

Mangone M, Manoharan AP, Thierry-Mieg D, Thierry-Mieg J, Han T, Mackowiak SD, Mis E, Zegar C, Gutwein MR, Khivansara V et al. 2010. The landscape of C. elegans 3'UTRs. *Science* **329**: 432-435.

Medley JC, Panzade G, Zinovyeva AY. 2021. microRNA strand selection: Unwinding the rules. *Wiley Interdiscip Rev RNA* **12**: e1627.

Mei Q, Li X, Meng Y, Wu Z, Guo M, Zhao Y, Fu X, Han W. 2012. A facile and specific assay for quantifying microRNA by an optimized RT-qPCR approach. *PloS one* **7**: e46890.

Meijer HA, Smith EM, Bushell M. 2014. Regulation of miRNA strand selection: follow the leader? *Biochem Soc Trans* **42**: 1135-1140.

Melendez A, Levined B. August 24, 2009. Autophagy in C. elegans. In *WormBook*, doi:1551-8507 (ed. TCeR Community). WormBook.

Murari E, Meadows D, Cuda N, Mangone M. 2024. A comprehensive analysis of 3'UTRs in Caenorhabditis elegans. *Nucleic Acids Res* doi:10.1093/nar/gkae543.

Nakanishi K. 2016. Anatomy of RISC: how do small RNAs and chaperones activate Argonaute proteins? *Wiley Interdiscip Rev RNA* **7**: 637-660.

Nakanishi K. 2024. When Argonaute takes out the ribonuclease sword. *J Biol Chem* **300**: 105499.

Pakravan K, Babashah S, Sadeghizadeh M, Mowla SJ, Mossahebi-Mohammadi M, Ataei F, Dana N, Javan M. 2017. MicroRNA-100 shuttled by mesenchymal stem cell-derived exosomes suppresses in vitro angiogenesis through modulating the mTOR/HIF-1alpha/VEGF signaling axis in breast cancer cells. *Cell Oncol (Dordr)* **40**: 457-470.

Pandit A, Jain V, Kumar N, Mukhopadhyay A. 2014. PHA-4/FOXA-regulated microRNA feed forward loops during Caenorhabditis elegans dietary restriction. *Aging (Albany NY)* **6**: 835-855.

Pare JM, Tahbaz N, Lopez-Orozco J, LaPointe P, Lasko P, Hobman TC. 2009. Hsp90 regulates the function of argonaute 2 and its recruitment to stress granules and P-bodies. *Mol Biol Cell* **20**: 3273-3284.

Ramani AK, Calarco JA, Pan Q, Mavandadi S, Wang Y, Nelson AC, Lee LJ, Morris Q, Blencowe BJ, Zhen M et al. 2011. Genome-wide analysis of alternative splicing in Caenorhabditis elegans. *Genome Res* **21**: 342-348.

Rane S, He M, Sayed D, Vashistha H, Malhotra A, Sadoshima J, Vatner DE, Vatner SF, Abdellatif M. 2009. Downregulation of miR-199a derepresses hypoxia-inducible factor-1alpha and Sirtuin 1 and recapitulates hypoxia preconditioning in cardiac myocytes. *Circ Res* **104**: 879-886.

Rhim J, Baek W, Seo Y, Kim JH. 2022. From Molecular Mechanisms to Therapeutics: Understanding MicroRNA-21 in Cancer. *Cells* **11**.

Rishik S, Hirsch P, Grandke F, Fehlmann T, Keller A. 2025. miRNATissueAtlas 2025: an update to the uniformly processed and annotated human and mouse non-coding RNA tissue atlas. *Nucleic Acids Res* **53**: D129-D137.

Schorr AL, Mejia AF, Miranda MY, Mangone M. 2023. An updated C. elegans nuclear body muscle transcriptome for studies in muscle formation and function. *Skelet Muscle* **13**: 4.

Shaw WR, Armisen J, Lehrbach NJ, Miska EA. 2010. The conserved miR-51 microRNA family is redundantly required for embryonic development and pharynx attachment in Caenorhabditis elegans. *Genetics* **185**: 897-905.

Smith A, Calley J, Mathur S, Qian HR, Wu H, Farmen M, Caiment F, Bushel PR, Li J, Fisher C et al. 2016. The Rat microRNA body atlas; Evaluation of the microRNA content of rat organs through deep sequencing and characterization of pancreas enriched miRNAs as biomarkers of pancreatic toxicity in the rat and dog. *BMC Genomics* **17**: 694.

Steber HS, Gallante C, O’Brien S, Chiu PL, Mangone M. 2019. The C. elegans 3’-UTRome v2 resource for studying mRNA cleavage and polyadenylation, 3’-UTR biology, and miRNA targeting. *Genome Research*.

Stein L, Mangone M, Schwarz E, Durbin R, Thierry-Mieg J, Spieth J, Sternberg P. 2001. WormBase: network access to the genome and biology of Caenorhabditis elegans (vol 29, pg 82, 2001). *Nucleic Acids Research* **29**: 1012.

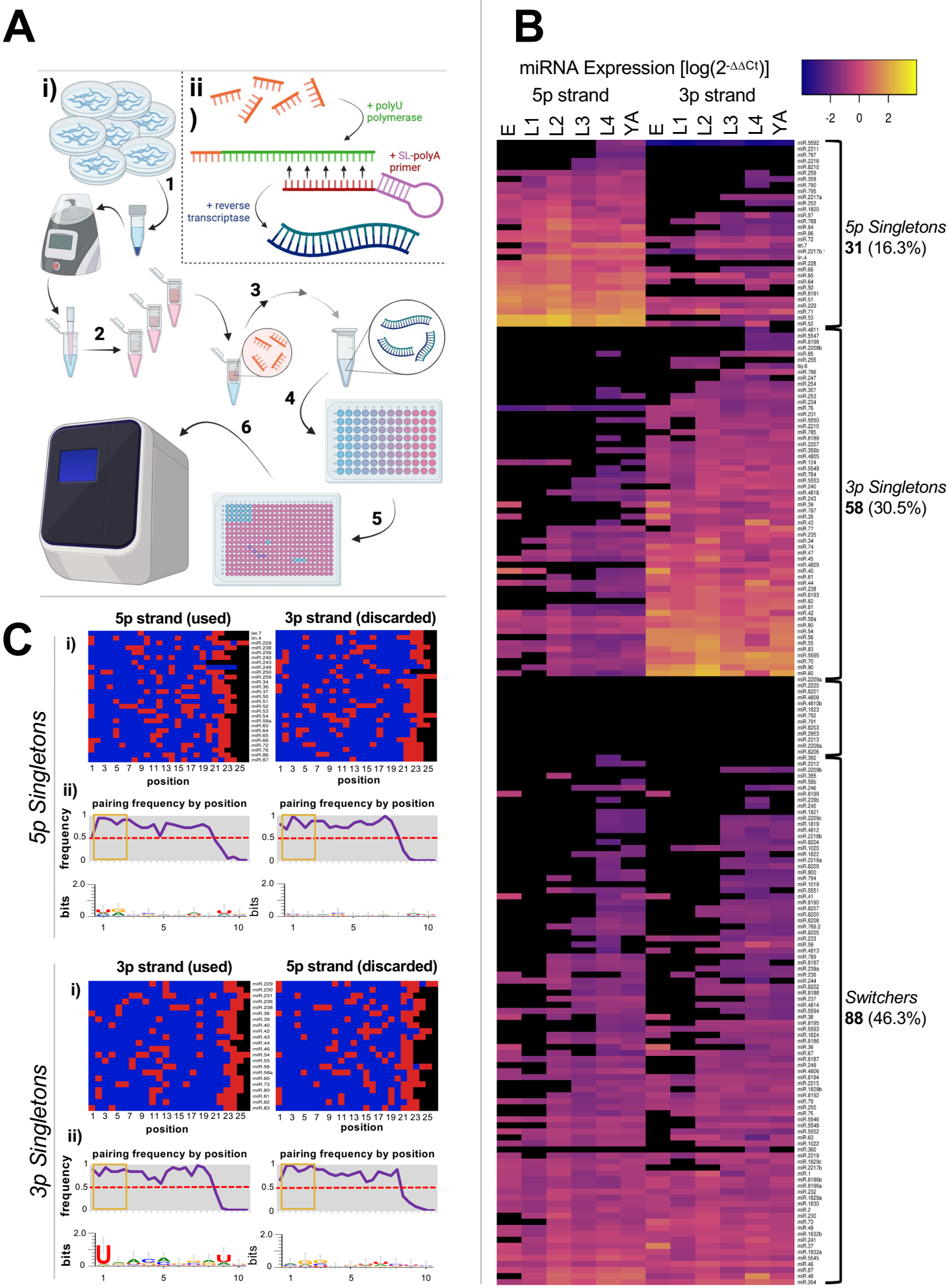
Sulston JE. 1983. Neuronal cell lineages in the nematode Caenorhabditis elegans. *Cold Spring Harb Symp Quant Biol* **48 Pt 2**: 443-452.

Wolter JM, Le HH, Linse A, Godlove VA, Nguyen TD, Kotagama K, Lynch A, Rawls A, Mangone M. 2017. Evolutionary patterns of metazoan microRNAs reveal targeting principles in the let-7 and miR-10 families. *Genome Res* **27**: 53-63.

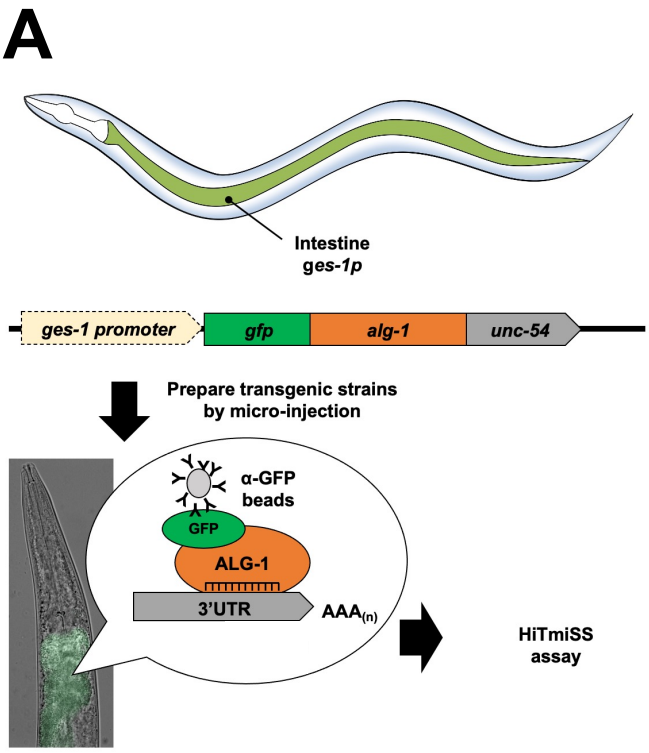
Zhang Q, Hrach H, Mangone M, Reiner DJ. 2022. Identifying the Caenorhabditis elegans vulval transcriptome. *G3 (Bethesda)* **12**.

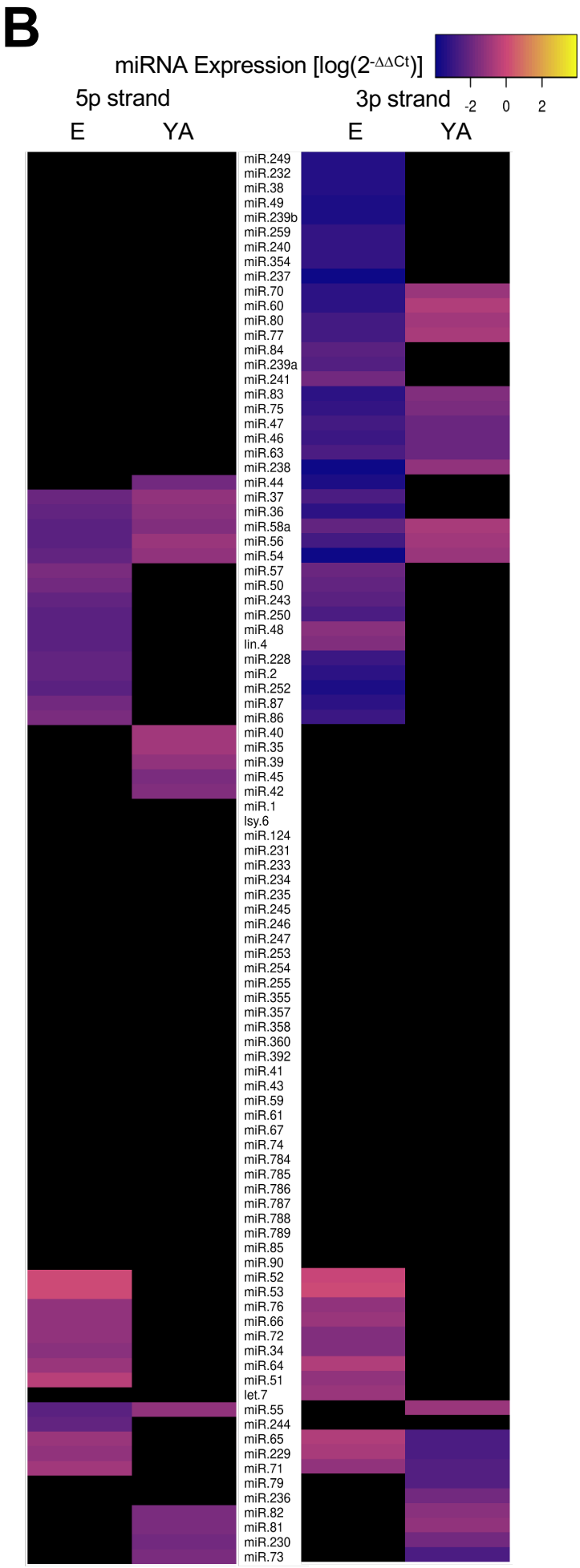
Zinovyeva AY, Veksler-Lublinsky I, Vashisht AA, Wohlschlegel JA, Ambros VR. 2015. Caenorhabditis elegans ALG-1 antimorphic mutations uncover functions for Argonaute in microRNA guide strand selection and passenger strand disposal. *Proc Natl Acad Sci U S A* **112**: E5271-5280.

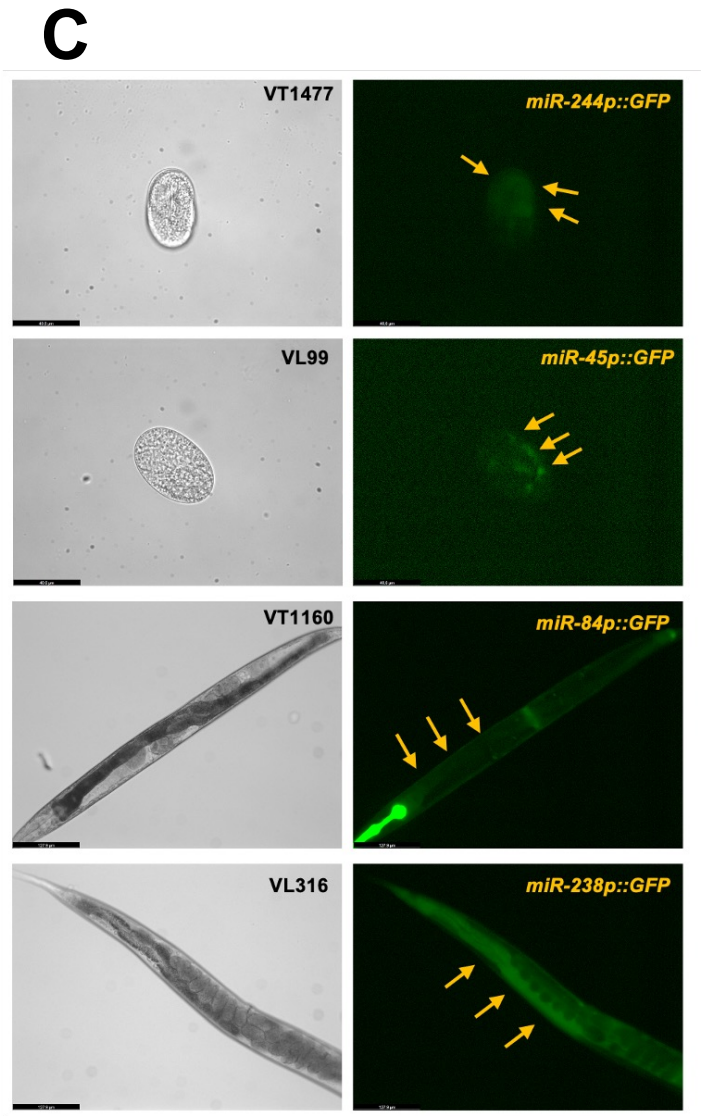
Main Figure 1



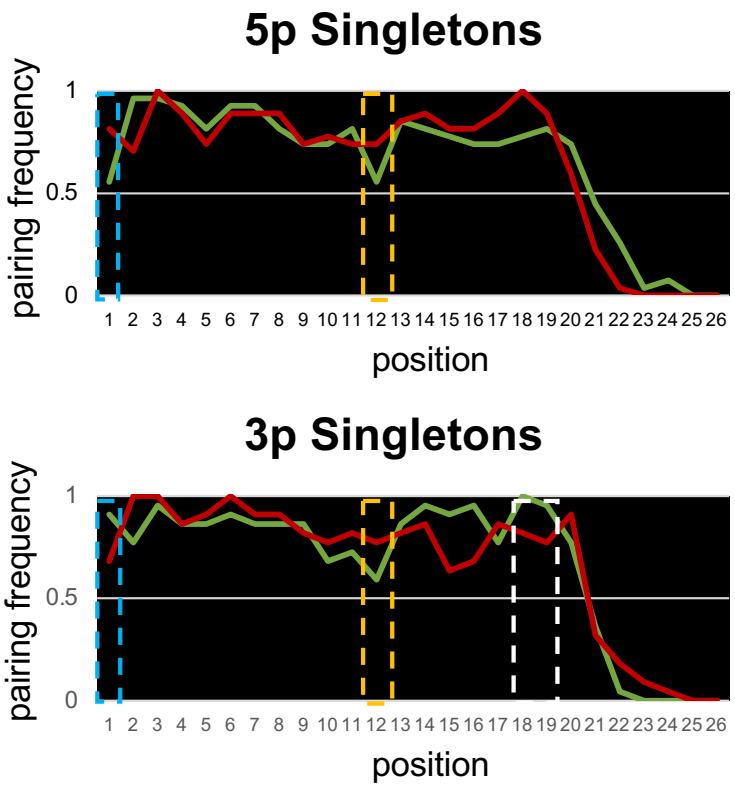
Main Figure 2



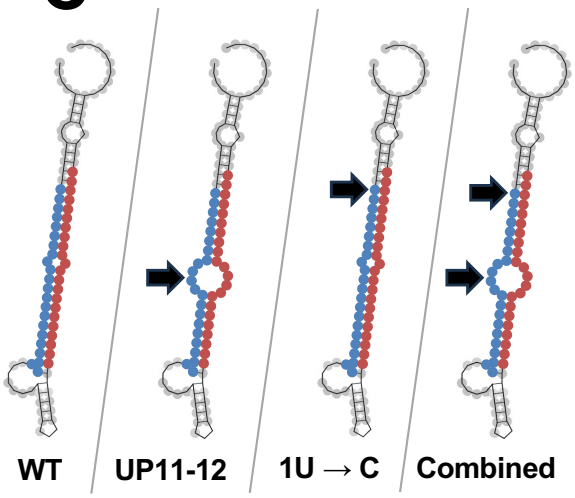




A

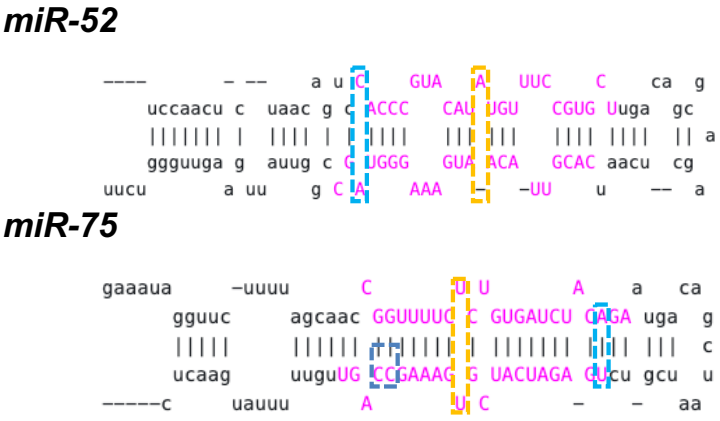


C

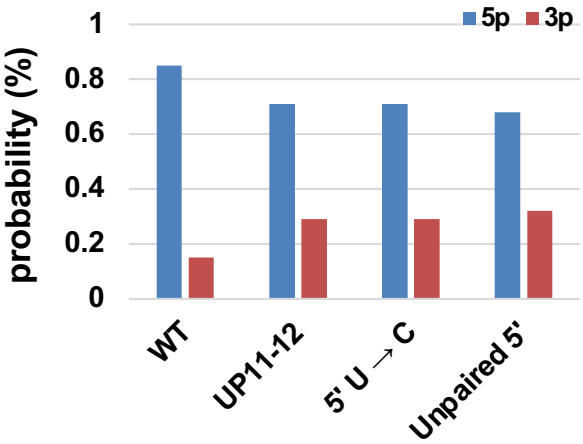


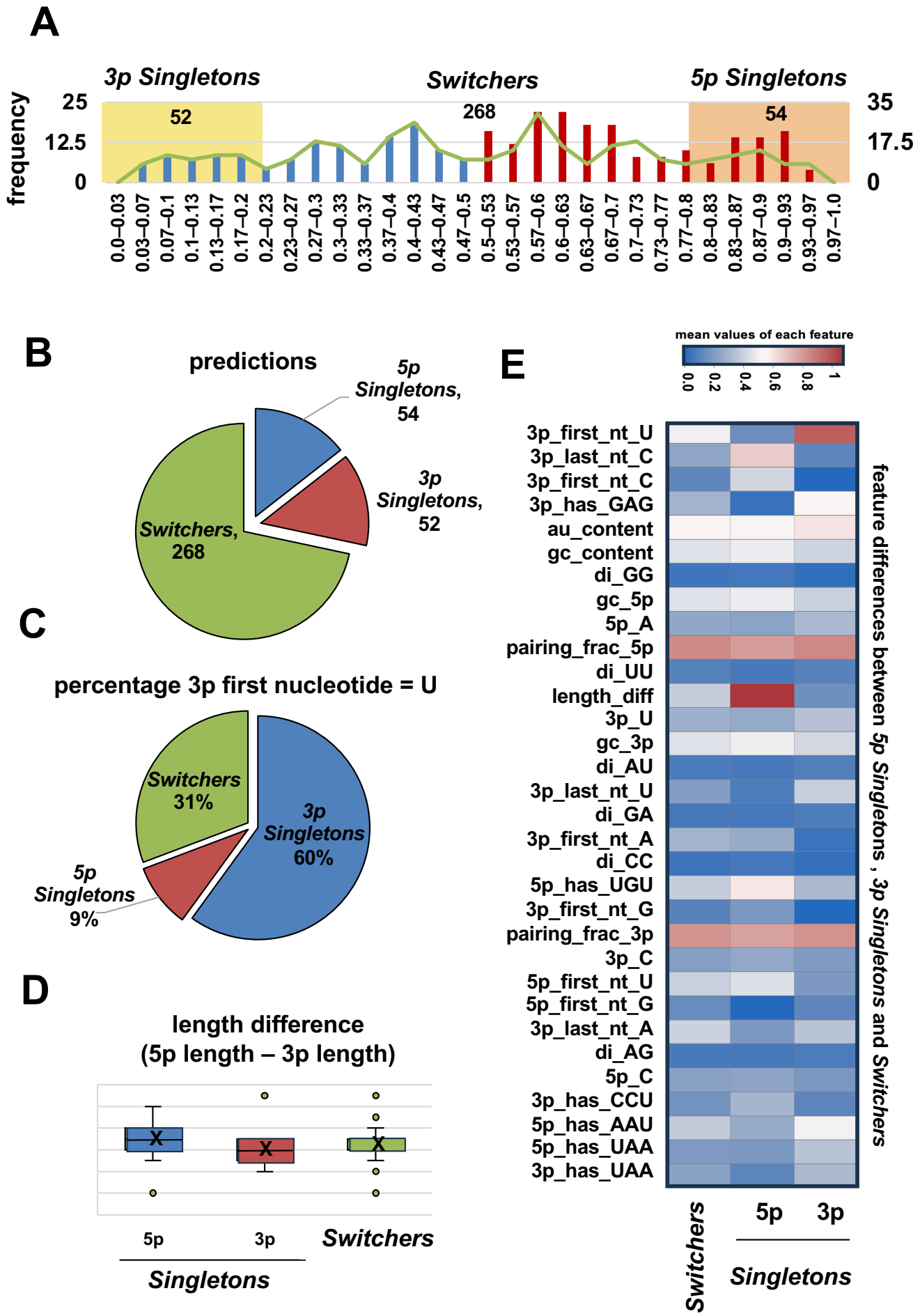
Variant	5p	3p
WT	0.85	0.15
UP11-12	0.71	0.29
5'U → C	0.71	0.29
Combined	0.68	0.32

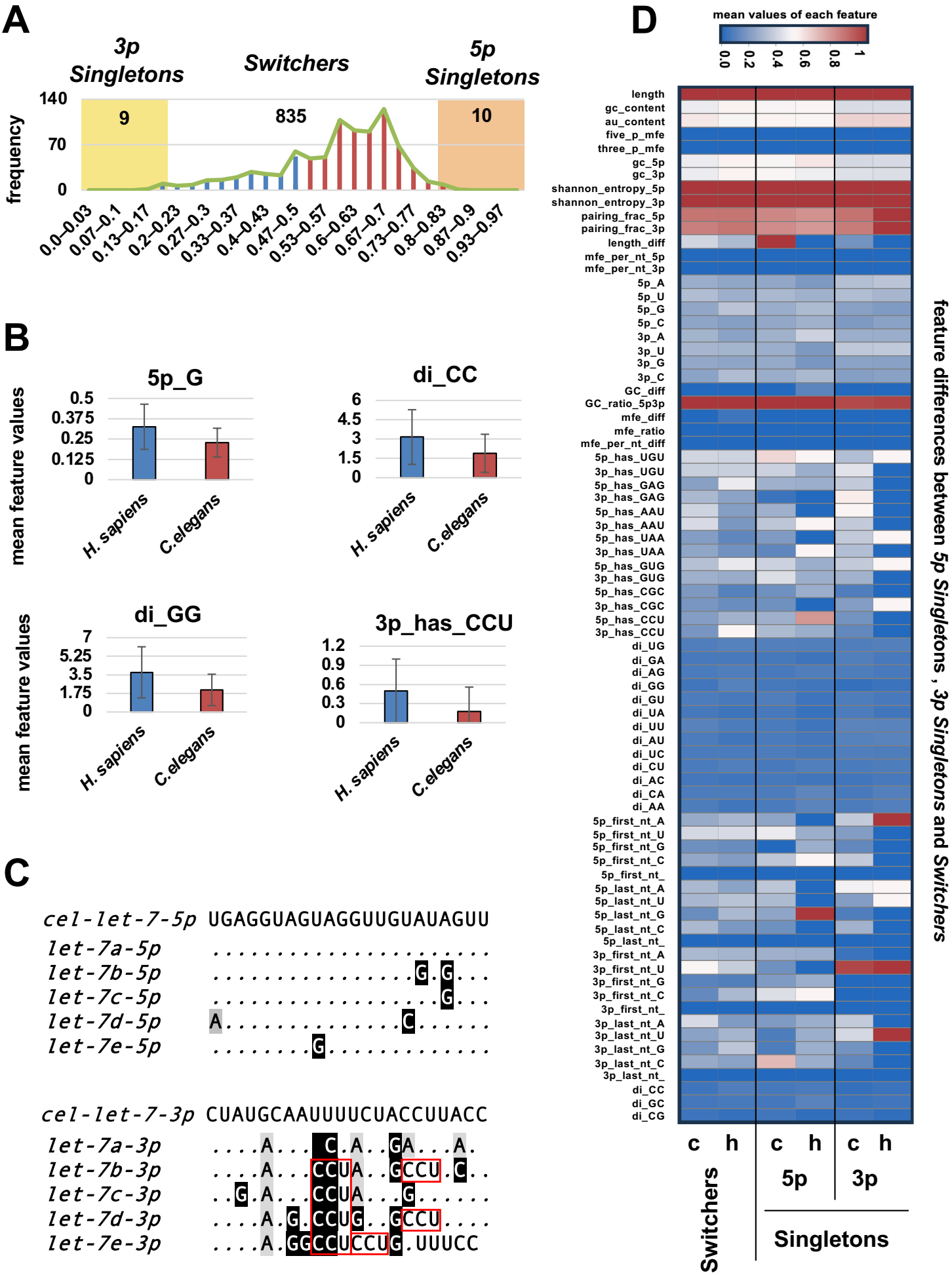
B

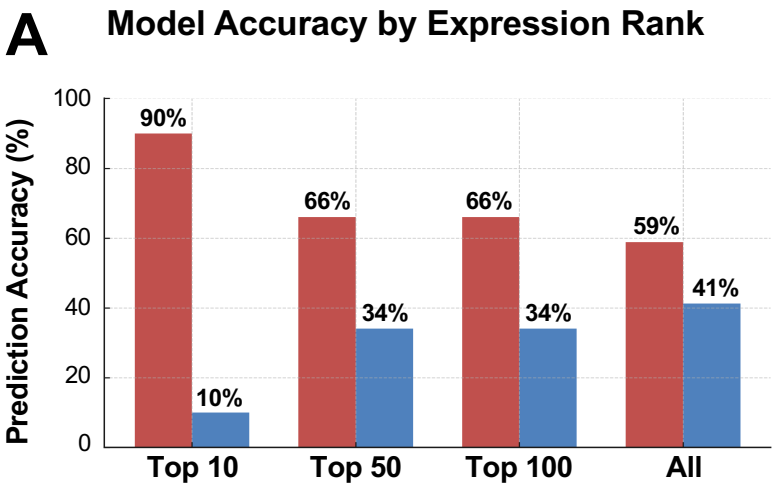


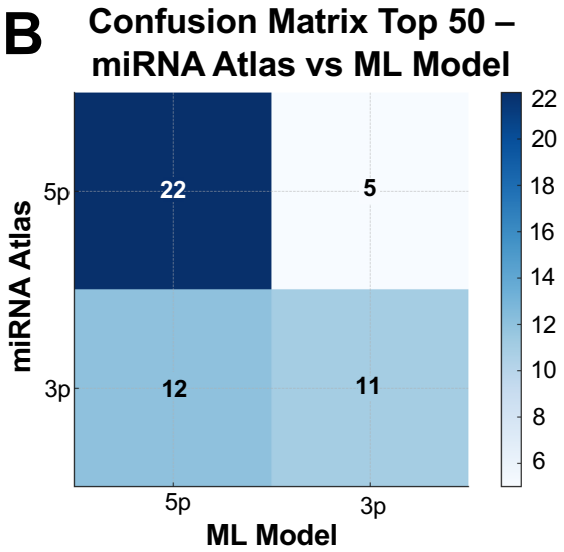
D

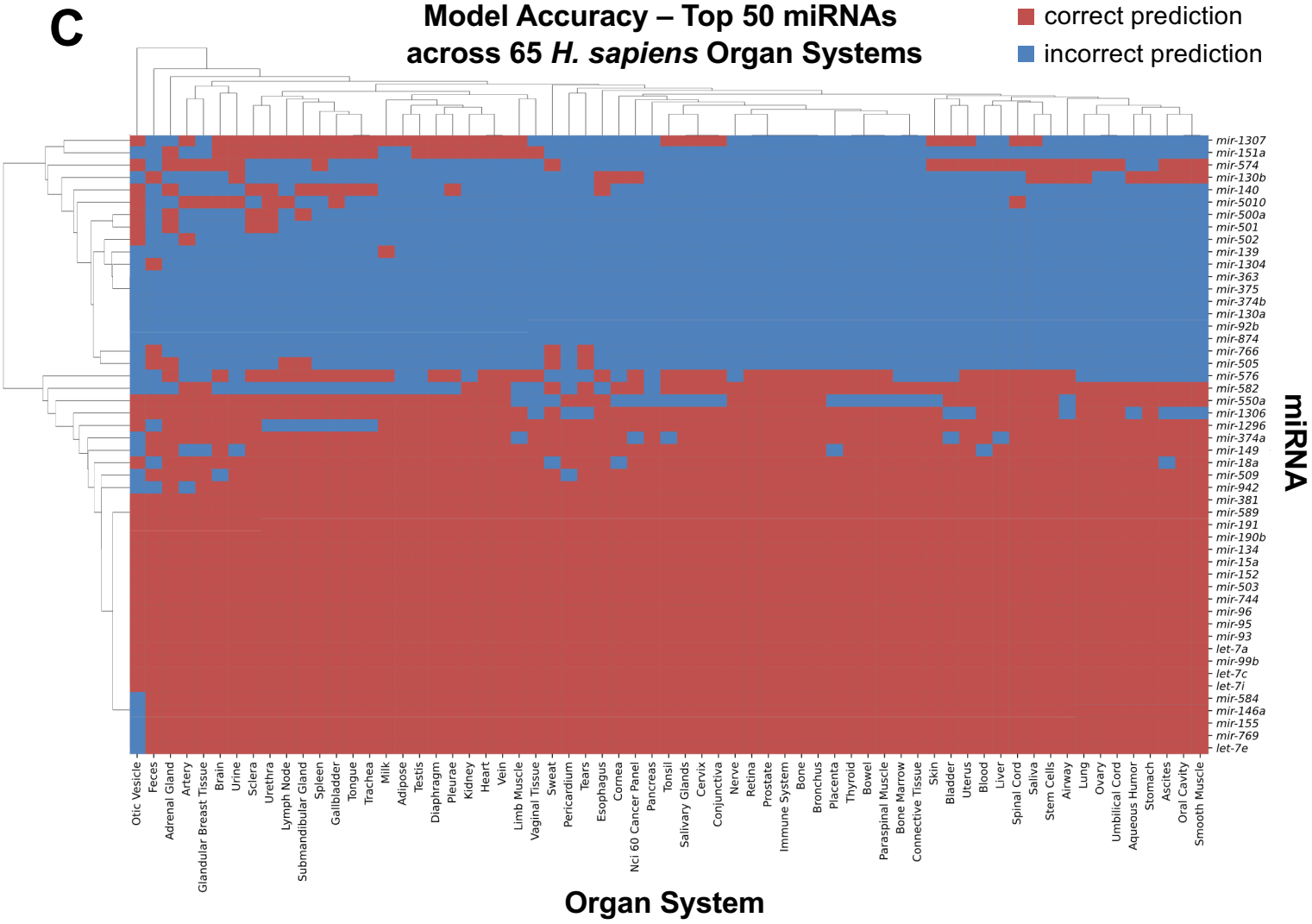



A
3p Singletons
Switchers
5p Singletons
frequency
25
12.5
0
35
17.5
0
52
268
54
0.0–0.03
0.03–0.07
0.07–0.1
0.1–0.13
0.13–0.17
0.17–0.2
0.2–0.23
0.23–0.27
0.27–0.3
0.3–0.33
0.33–0.37
0.37–0.4
0.4–0.43
0.43–0.47
0.47–0.5
0.5–0.53
0.53–0.57
0.57–0.6
0.6–0.63
0.63–0.67
0.67–0.7
0.7–0.73
0.73–0.77
0.77–0.8
0.8–0.83
0.83–0.87
0.87–0.9
0.9–0.93
0.93–0.97
0.97–1.0
B
predictions
5p Singletons, 54
3p Singletons, 52
Switchers, 268
C
percentage 3p first nucleotide = U
Switchers 31%
3p Singletons 60%
5p Singletons 9%
D
length difference
(5p length – 3p length)
5p
3p
Switchers
Singletons
E
mean values of each feature
0.0
0.2
0.4
0.6
0.8
1
3p_first_nt_U
3p_last_nt_C
3p_first_nt_C
3p_has_GAG
au_content
gc_content
di_GG
gc_5p
5p_A
pairing_frac_5p
di_UU
length_diff
3p_U
gc_3p
di_AU
3p_last_nt_U
di_GA
3p_first_nt_A
di_CC
5p_has_UGU
3p_first_nt_G
pairing_frac_3p
3p_C
5p_first_nt_U
5p_first_nt_G
3p_last_nt_A
di_AG
5p_C
3p_has_CCU
5p_has_AAU
5p_has_UAA
3p_has_UAA
feature differences between 5p Singletons , 3p Singletons and Switchers
Switchers
5p
3p
Singletons


A
3p Singletons
Switchers
5p Singletons
frequency
140
70
0
9
835
10
0.0–0.03
0.07–0.1
0.13–0.17
0.2–0.23
0.27–0.3
0.33–0.37
0.4–0.43
0.47–0.5
0.53–0.57
0.6–0.63
0.67–0.7
0.73–0.77
0.8–0.83
0.87–0.9
0.93–0.97
B
mean feature values
5p_G
0.5
0.375
0.25
0.125
0
di_CC
6
4.5
3
1.5
0
di_GG
7
5.25
3.5
1.75
0
3p_has_CCU
1.2
0.9
0.6
0.3
0
H. sapiens
C.elegans
C
cel-let-7-5p UGAGGUAGUAGGUUGUAUAGUU
let-7a-5p
let-7b-5p
let-7c-5p
let-7d-5p
let-7e-5p
cel-let-7-3p CUAUGCAAUUUUCUACCUUACC
let-7a-3p
let-7b-3p
let-7c-3p
let-7d-3p
let-7e-3p
D
mean values of each feature
0.0
0.2
0.4
0.6
0.8
1
length
gc_content
au_content
five_p_mfe
three_p_mfe
gc_5p
gc_3p
shannon_entropy_5p
shannon_entropy_3p
pairing_frac_5p
pairing_frac_3p
length_diff
mfe_per_nt_5p
mfe_per_nt_3p
5p_A
5p_U
5p_G
5p_C
3p_A
3p_U
3p_G
3p_C
GC_diff
GC_ratio_5p3p
mfe_diff
mfe_ratio
mfe_per_nt_diff
5p_has_UGU
3p_has_UGU
5p_has_GAG
3p_has_GAG
5p_has_AAU
3p_has_AAU
5p_has_UAA
3p_has_UAA
5p_has_GUG
3p_has_GUG
5p_has_CGC
3p_has_CGC
5p_has_CCU
3p_has_CCU
di_UG
di_GA
di_AG
di_GG
di_GU
di_UA
di_UU
di_AU
di_UC
di_CU
di_AC
di_CA
di_AA
5p_first_nt_A
5p_first_nt_U
5p_first_nt_G
5p_first_nt_C
5p_first_nt_
5p_last_nt_A
5p_last_nt_U
5p_last_nt_G
5p_last_nt_C
5p_last_nt_
3p_first_nt_A
3p_first_nt_U
3p_first_nt_G
3p_first_nt_C
3p_first_nt_
3p_last_nt_A
3p_last_nt_U
3p_last_nt_G
3p_last_nt_C
3p_last_nt_
di_CC
di_GC
di_CG
c h c h c h
5p
3p
Switchers
Singletons
feature differences between 5p Singletons , 3p Singletons and Switchers


A
Model Accuracy by Expression Rank
Prediction Accuracy (%)
100
80
60
40
20
0
90%
10%
66%
34%
66%
34%
59%
41%
Top 10
Top 50
Top 100
All


B
Confusion Matrix Top 50 – miRNA Atlas vs ML Model
miRNA Atlas
5p
3p
22
5
12
11
ML Model
5p
3p
22
20
18
16
14
12
10
8
6


C
Model Accuracy – Top 50 miRNAs across 65 H. sapiens Organ Systems
correct prediction
incorrect prediction
miRNA
mir-1307
mir-151a
mir-574
mir-130b
mir-140
mir-5010
mir-500a
mir-501
mir-502
mir-139
mir-1304
mir-363
mir-375
mir-374b
mir-130a
mir-92b
mir-874
mir-766
mir-505
mir-576
mir-582
mir-550a
mir-1306
mir-1296
mir-374a
mir-149
mir-18a
mir-509
mir-942
mir-381
mir-589
mir-191
mir-190b
mir-134
mir-15a
mir-152
mir-503
mir-744
mir-96
mir-95
mir-93
let-7a
mir-99b
let-7c
let-7i
mir-584
mir-146a
mir-155
mir-769
let-7e
Otic Vesicle
Feces
Adrenal Gland
Artery
Glandular Breast Tissue
Brain
Urine
Sclera
Urethra
Lymph Node
Submandibular Gland
Spleen
Gallbladder
Tongue
Trachea
Milk
Adipose
Testis
Diaphragm
Pleurae
Kidney
Heart
Vein
Limb Muscle
Vaginal Tissue
Sweat
Pericardium
Tears
Esophagus
Cornea
Nci 60 Cancer Panel
Pancreas
Tonsil
Salivary Glands
Cervix
Conjunctiva
Nerve
Retina
Prostate
Immune System
Bone
Bronchus
Placenta
Thyroid
Bowel
Paraspinal Muscle
Bone Marrow
Connective Tissue
Skin
Bladder
Uterus
Blood
Liver
Spinal Cord
Saliva
Stem Cells
Airway
Lung
Ovary
Umbilical Cord
Aqueous Humor
Stomach
Ascites
Oral Cavity
Smooth Muscle
Organ System

Additional File 1 for:

Exploring MicroRNA Strand Selection: Insights into Gene Regulation and Development

Dalton Meadows[1,2], Halee Hargis[1,2], Amanda Ellis[1,2] and Marco Mangone[2*]

[1)] The Biodesign Institute at Arizona State University 1001 S McAllister Ave, Tempe, AZ

[2)] School of Life Sciences, Arizona State University, 427 E Tyler Mall, Tempe, AZ

* To whom correspondence should be addressed. Tel: (480) 965-7957; Email: mangone@asu.edu

This PDF includes:

Supplemental Figs. S1-S6

Table of Contents

Additional File 1

Fig. S1: Overview of miRNA Strand Usage Prediction Model.....3

Fig. S2: Overview of Model Performance.....4

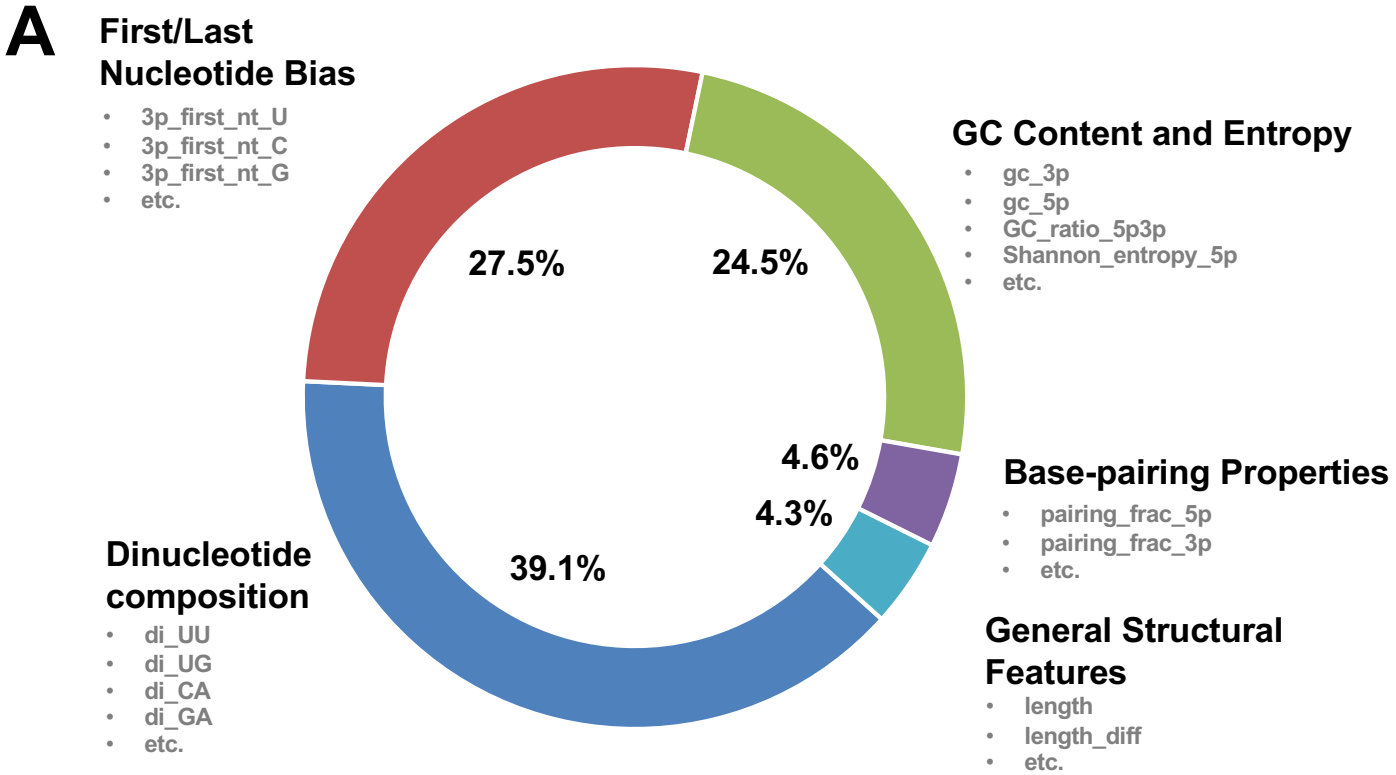
Fig. S3: Strand Usage per Category in Representative miRNAs.....5

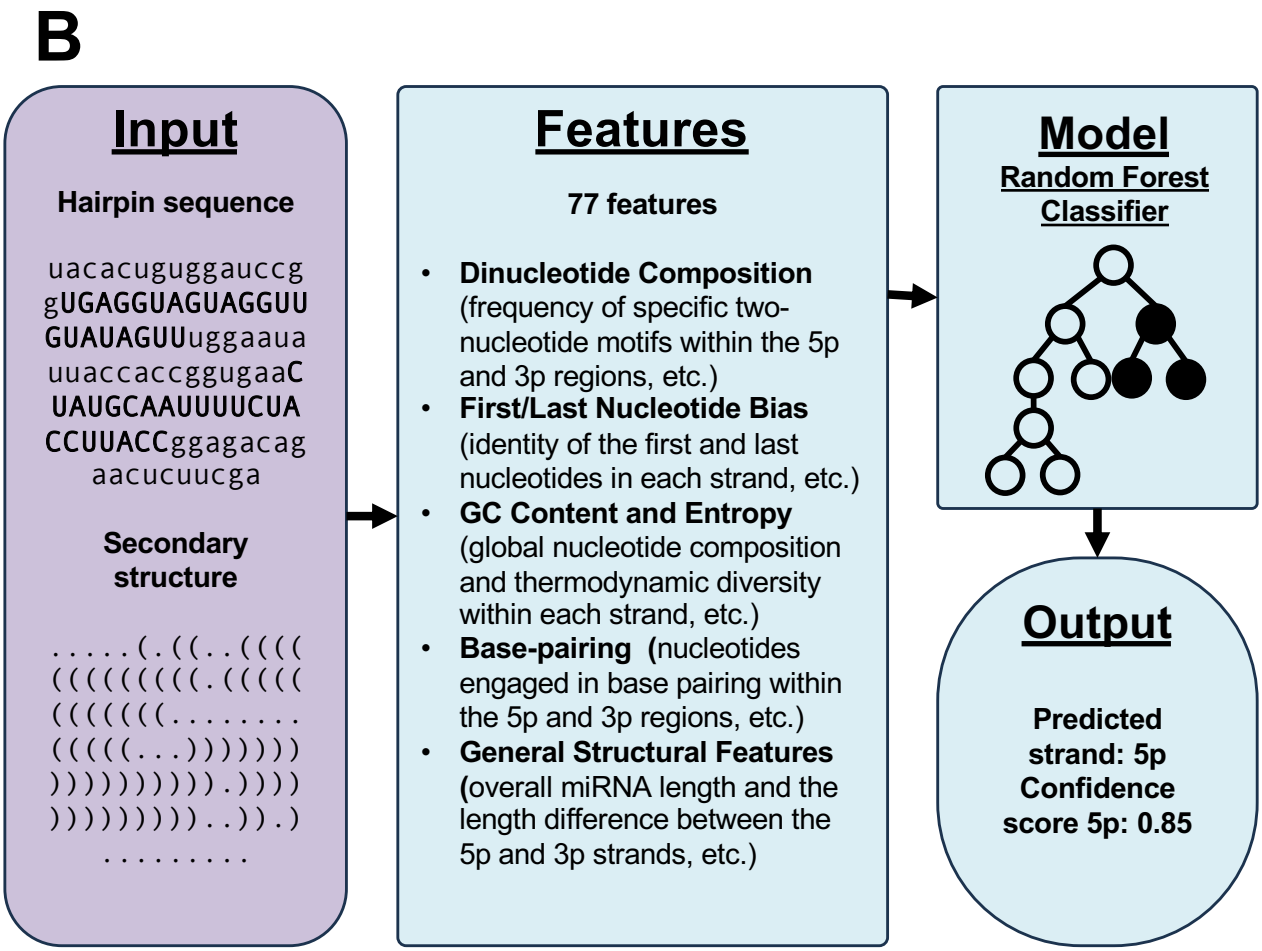
Fig. S4: HiTmiSS Comparison and Intestine Tissue *Switchers*.....6

Fig. S5: Testing the Target Stabilization Hypothesis.....7

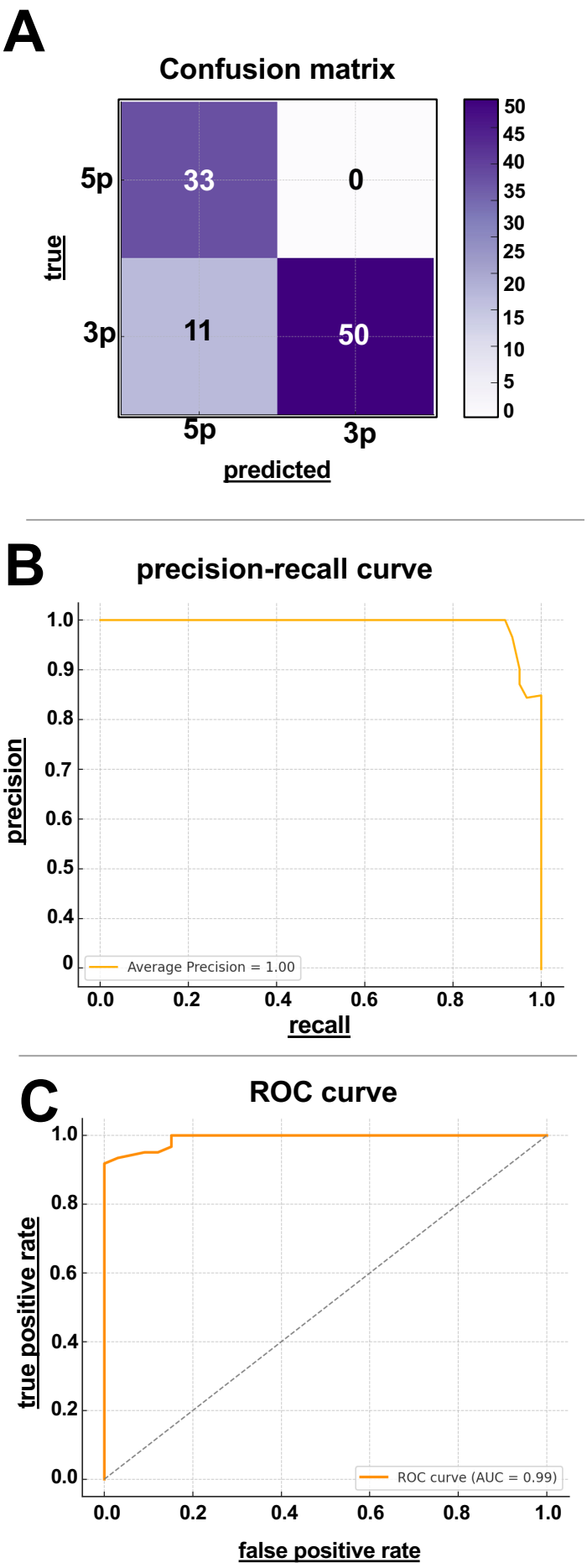
Fig. S6: Predicted Strand Usage in *5p* and *3p Singletons*.....8

Supplemental Figure S1

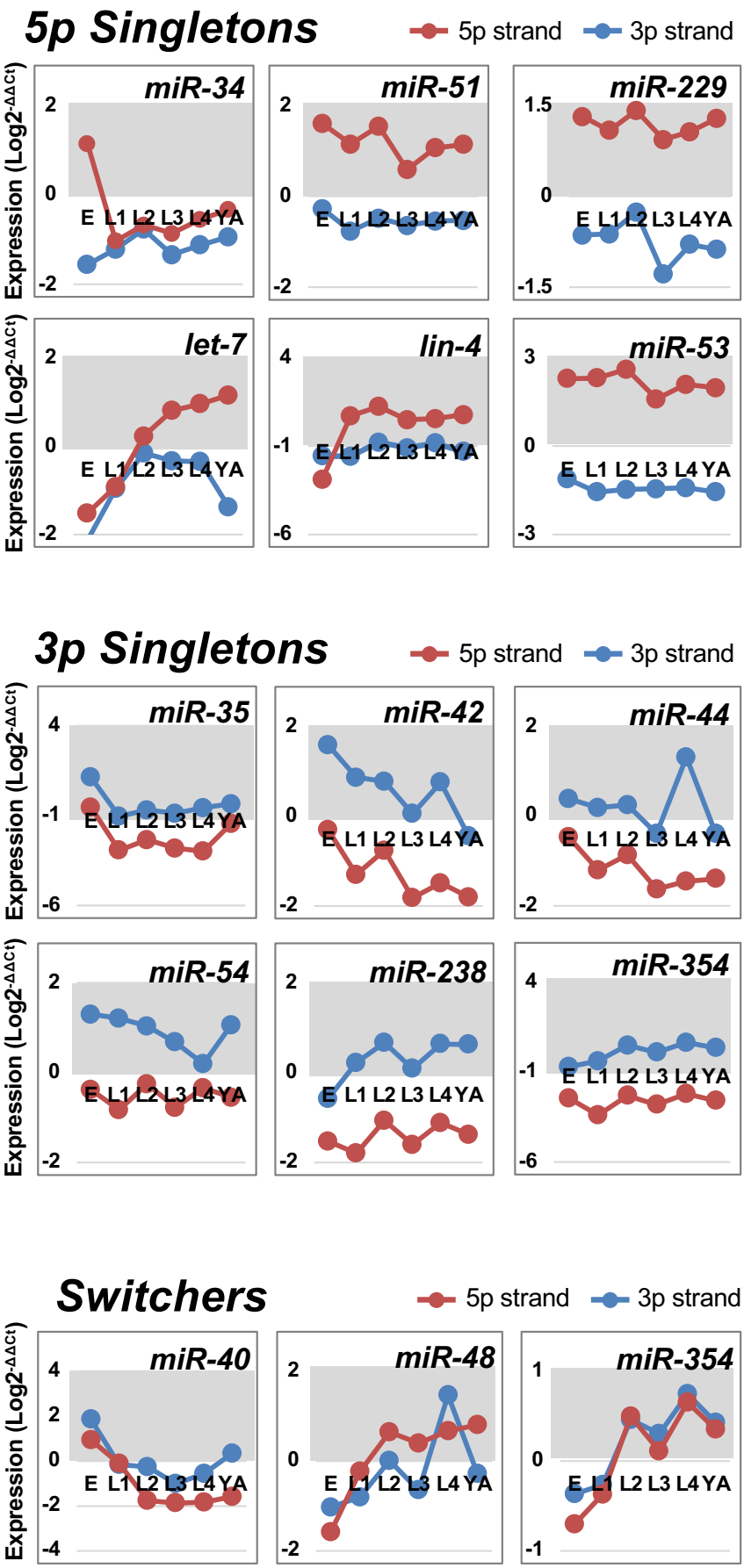




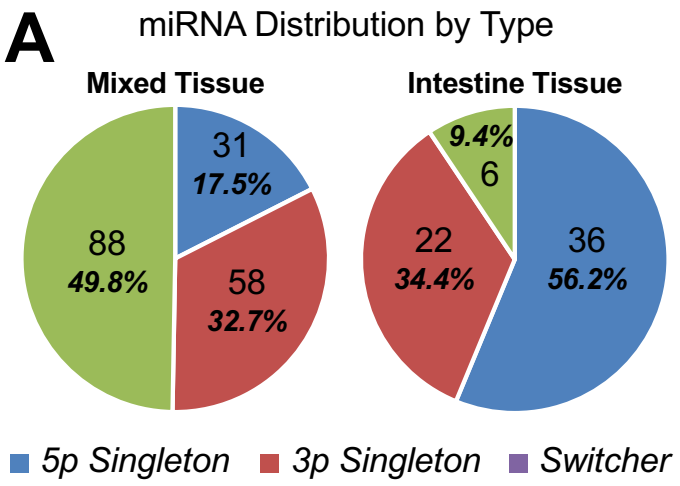
Supplemental Figure S1: Overview of the miRNA Strand Usage Prediction Model. **(A)** Weight assigned to each feature of the model. **(B)** Flowchart illustrating the computational pipeline developed to predict 5p or 3p strand usage in miRNA precursors using machine learning. Sample *let-7* input and output shown.



Supplemental Figure S2: Overview of Model Performance. **(A)** Confusion matrix of the strand usage prediction model. Summarizes the classification performance of the Random Forest model trained on *C. elegans* miRNA data, showing the number of correctly (33 and 50) and incorrectly (0 and 11) predicted 5p and 3p strands. **(B)** Precision-recall curve illustrating the relationship between precision (positive predictive value) and recall (sensitivity) for classifying 5p versus 3p strand usage in miRNA precursors using the Random Forest model. Each point on the curve represents a different decision threshold. The shape of the curve reflects the model's ability to maintain high precision while recovering true positives across a range of thresholds. **(C)** Receiver Operating Characteristic (ROC) curve of the strand usage prediction model. The ROC curve shows the trade-off between true positive rate (sensitivity) and false positive rate (specificity) for predicting 5p versus 3p strand usage using a Random Forest model trained on *C. elegans* miRNA data. Each point represents a different classification threshold. The area under the curve (AUC) provides a summary measure of the model's overall performance, with higher values indicating stronger discrimination between 5p and 3p strands. The curve demonstrates reliable strand separation across a range of thresholds.



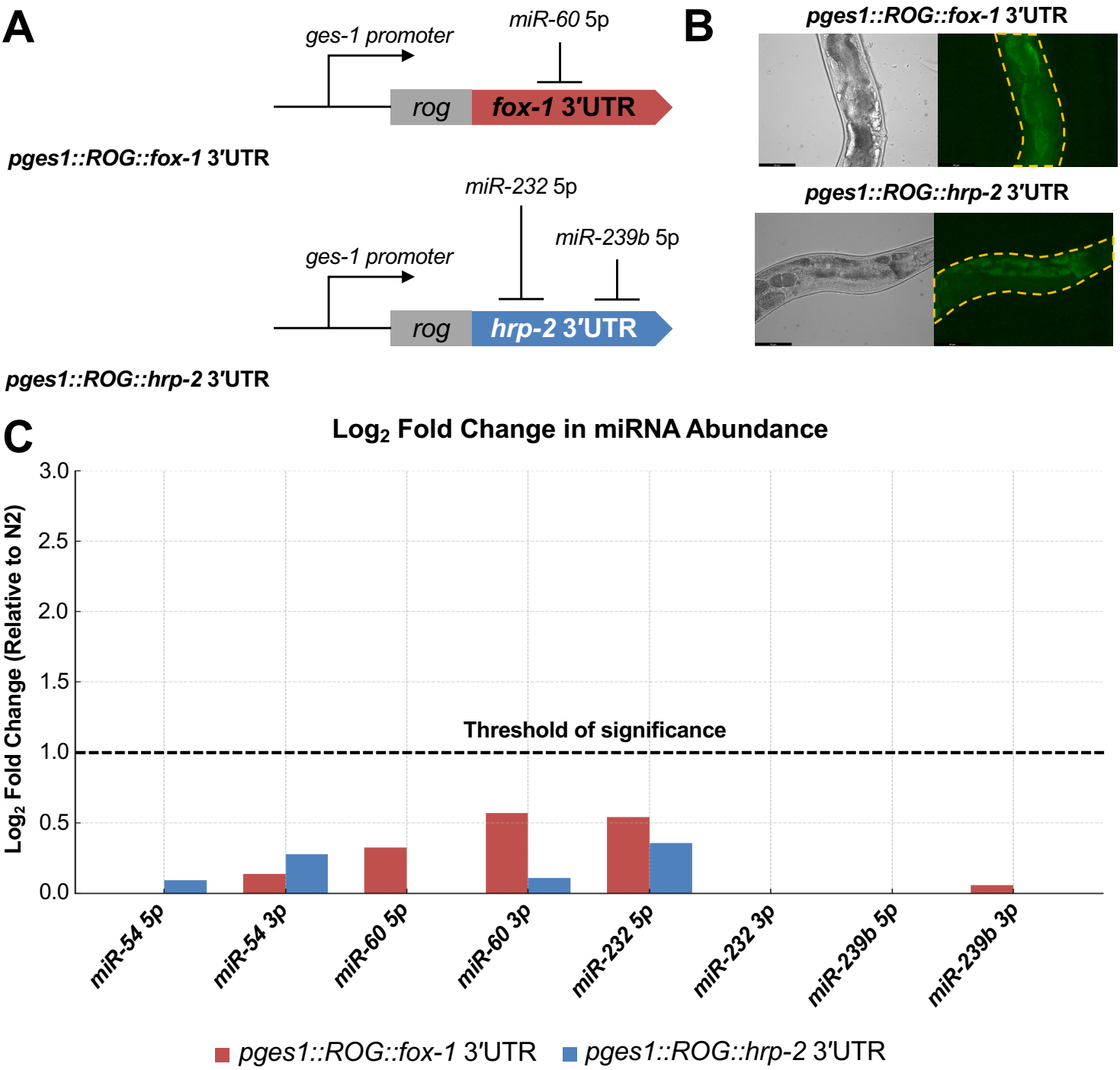
Supplemental Figure S3: Strand Usage per Category in Representative miRNAs. Normalized mean expression values (n=3) for the 5p strand (red) and 3p strand (blue) of several representative miRNAs that fall into one of our categories, either *5p Singleton*, *3p Singleton*, or *Switcher*. Expression is demonstrated to fluctuate across developmental time for each of these miRNAs in mixed tissue, following their expression from embryo stage (E) through the four larval stages (L1-L4) and in young adult stage (YA). Within each *Singleton* category we find well characterized canonical regulators of development such as *let-7*, *lin-4*, *miR-35* and *miR-53*. Within the *Switcher* category the trends are more diverse, with some miRNAs like *miR-40* altering which strand is used after embryo stage, while others like *miR-48* display strand switching later in development. Still others like *miR-354* show relatively equal strand usage throughout all developmental stages. Perhaps this diversity in strand usage trend among *Switchers* in our whole organism data can be explained by tissue-specific regulation, with one strand preferred in some tissues and the opposite strand preferred in other tissues.



B

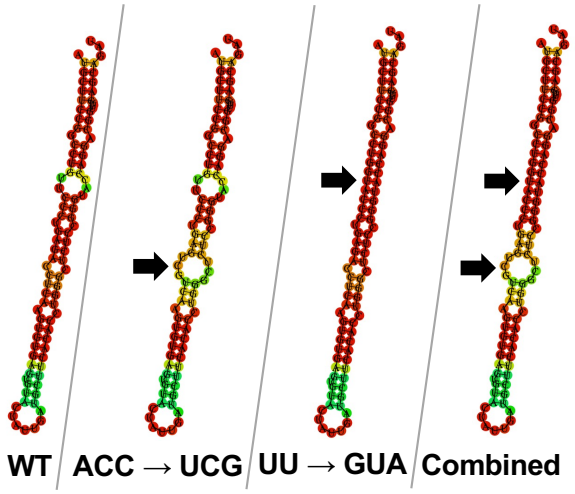
Family	Family Members	Function in the Intestine
miR-44	*miR-44*	heterochronic timing, development
miR-51	*miR-54, miR-56, miR-65*	polarity, hypoxia
miR-58	*miR-58a*	stress response, dauer entry
miR-71	*miR-71*	starvation, longevity

Supplemental Figure S4: HiTmiSS Comparison and Intestine Tissue *Switchers*. (A) Comparison of the distribution of *5p Singletons*, *3p Singletons*, and *Switchers* between HiTmiSS analysis results in mixed tissue (left) and intestine tissue (right). Results expressed as a percentage of total detected miRNAs per analysis. *Singleton* strand preference shifted to favor *5p Singletons* and overall abundance of *Switchers* steeply declined, both of which suggest that strand selection is regulated at the level of individual tissues. **(B)** Breakdown of *Switchers* detected in the intestine tissue by miRNA family and function. *Switcher* in this case was defined as any miRNA whose strand preference changed from embryo to young adult stage, either 5p to 3p or vice versa. Though few in number, each *Switcher* regulates target 3′UTRs of genes directly involved in the function of the intestine, further reinforcing the importance of the tissue-specific regulation of strand selection.



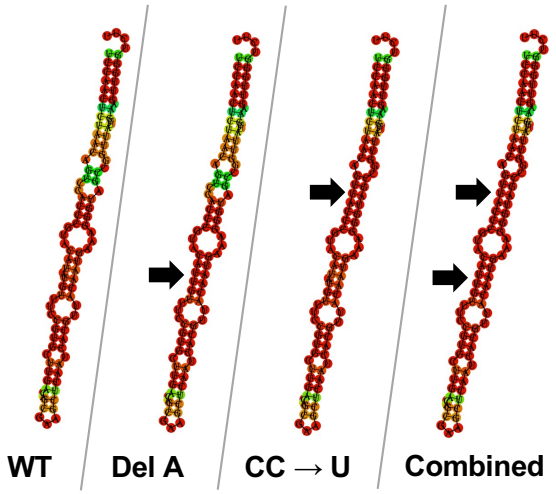
Supplemental Figure S5: Testing the Target Stabilization Hypothesis. **(A)** Constructs used to overexpress GFP under the control of the *fox-1* 3′UTR (top) or *hrp-2* 3′UTR (bottom) using the promoter for the intestine-specific gene *ges-1.* miRNA strands which bind each target 3′UTR (shown above) are typically not expressed in wildtype N2 *C. elegans*. (B) Representative brightfield (left) and GFP (right) images of C. elegans strains expressing these constructs, demonstrating fluorescence in the intestine. (C) Change in abundance of each miRNA strand, expressed as $\log_2$ of the difference in mean expression value as per our HiTmiSS assay. None of the tested miRNA strands were significantly ($\log_2 > 1$) upregulated when overexpressing either test 3′UTR. These data suggest that overexpression of targets for a miRNA strand does not lead to retention of that strand, but rather strand selection occurs independent of the presence or absence of target 3′UTRs.

lin-4



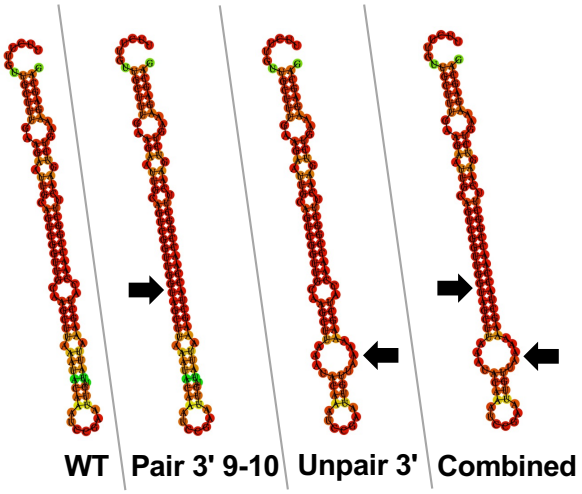
Variant	5p	3p
WT	0.81	0.19
ACC → UCG	0.74	0.26
UU → GUA	0.77	0.23
Combined	0.76	0.24

miR-52



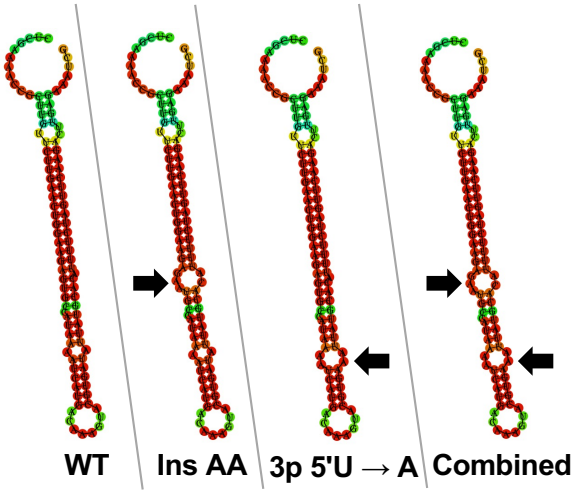
Variant	5p	3p
WT	0.89	0.11
Del A	0.83	0.17
CC → U	0.90	0.10
Combined	0.88	0.12

miR-75



Variant	5p	3p
WT	0.36	0.64
Pair 3' 9-10	0.30	0.70
Unpair 3'	0.60	0.40
Combined	0.57	0.43

miR-60



Variant	5p	3p
WT	0.41	0.59
Ins AA	0.46	0.54
3p 5'U → A	0.46	0.54
Combined	0.52	0.48

Supplemental Figure S6: Predicted Strand Usage in *5p* and *3p Singletons*. Representative, well-characterized *5p Singletons* (top) and *3p Singletons* (bottom) were selected and their strand preference predicted by our ML model (WT, far left). Each underwent two *in silico* site-directed mutations, one in the bulge region (middle left) and one in the pairing/identity of the first nucleotide (middle right). Finally, a combination mutation of both is shown (far right). The confidence of prediction is displayed as a decimal, both *5p Singleton* confidence (5p) or *3p Singleton* confidence (3p). These mutations consistently altered predicted strand confidence, with a more pronounced effect in *3p Singletons*, supporting the trends within our wet bench data.